STAT 530 Exam 2 - Due by 5:30pm Tuesday, December 6th

- The exam should be turned into me, or the secretary in room 216; it should not be left in my mailbox.
- You may use your notes, any text or reference book, the course web page, and any combination of R, SAS, and SPSS
- You may not discuss the problems with anyone (especially your fellow students or other instructors) except me. I am happy to give computer advice.
- You must turn in the code/commands used to generate the output.
- The exam is worth 80 points. Question 1 is worth 10 points and questions 2-6 are worth 14 points each. Notice that students taking the course for graduate credit must answer one of the two additional questions.

The raw data set used for all of the following questions can be found at:
http://www.stat.sc.edu/~habing/courses/data/finbears.txt

It is an expanded version of a data set we looked at earlier this semester and is based on a data set originally described in "Reader's Digest" (April, 1979) and "Sports Afield", (September, 1981). It consists of several measurements for bears that were captured, measured, and released. (The full data set actually caught several of the bears multiple times over a period of years.)

The variables in the data set are:

<table>
<thead>
<tr>
<th>Name</th>
<th>Sex</th>
<th>Age</th>
<th>Maturity</th>
<th>Weight</th>
<th>Length</th>
<th>ChestG</th>
<th>HeadL</th>
<th>HeadW</th>
<th>NeckG</th>
<th>sChestG</th>
<th>sHeadL</th>
<th>sHeadW</th>
<th>sNeckG</th>
<th>Grp</th>
</tr>
</thead>
<tbody>
<tr>
<td>Name – Name of the Bear</td>
<td>Sex – M=Male, F=Female</td>
<td>Age – Estimated Age in Months</td>
<td>Maturity – mature if more than 18 months, otherwise young</td>
<td>Weight – Weight in Pounds</td>
<td>Length – Body Length in Inches</td>
<td>ChestG – Girth of Chest in Inches</td>
<td>HeadL – Length of Head in Inches</td>
<td>HeadW – Width of Head in Inches</td>
<td>NeckG – Girth of Neck in Inches</td>
<td>sChestG – Standardized ChestG divided by Length</td>
<td>sHeadL – Standardized HeadL divided by Length</td>
<td>sHeadW – Standardized HeadW divided by Length</td>
<td>sNeckG – Standardized NeckG divided by Length</td>
<td>Grp – First letter indicates sex, second is Y for young and O for mature</td>
</tr>
</tbody>
</table>

The observations are currently ordered by name.

If it is read into R as fbears it may be useful to use attach(fbears) so that you can refer to each column merely by its name.

Also recall that:
cbind will produce a matrix of the selected columns, as in cbind(Length,NeckG,sNeckG)
as.matrix can do this as well, as in as.matrix(fbears[,c(6,10,14)])
as.character will convert a factor to a string, as in as.character(Name)
== will let you indicate a subset of matrix, as in fbears[Grp=="FO",]
1) Each part of this question can be answered simply by providing the appropriate name.

a) What multivariate statistical method would both give an idea of how many underlying latent variables are needed to model the quantitative variables in this data set as well as to describe what the relationships are between the latent and observed variables?

b) Which method from the first half of the course gives the same result as classical multidimensional scaling using Euclidean distance (be specific).

c) What multivariate method is used to test if a specific path diagram describes the underlying relationship between the variables in this data set.

d) If bears took multiple choice tests, what group of multivariate statistical methods would be useful for analyzing their test results.

e) What type of distance is very closely related to multivariate normality, the chi-square plot for checking multivariate normality, and Hotelling’s T-square?

2) It is desired to see if the four standardized sChestG, sHeadL, sHeadW, sNeckG, along with the overall Length are able to separate the bears based on sex and maturity group.

Perform a linear discriminant analysis to find the combinations of these five variables that best separate the bears into the four groups. Provide an appropriate measure of how well the linear discriminant functions do at separating the four groups overall, and indicate which of the groups seem to be identified well and which seem to be identified poorly.

One piece of the output produced in such an analysis is the set of posterior probabilities. Give the estimated posterior probabilities for the first bear Adam and briefly say what these probabilities indicate (imagine you are explaining to someone who knows very little about discriminant analysis and doesn’t want much detail). What assumptions need to be true in order for these probabilities to be accurate? (You do not need to check the assumptions).

3) In many instances it is desirable to find groups of individuals that are similar, whether it is to have control over variation in a designed experiment or for casting purposes in movies.

Find a group of four bears such that none of them differ very much from any of the others in the group in terms of any of the four rescaled and standardized variables sChestG, sHeadL, sHeadW, and sNeckG.

Make a table showing how those four bears compare on each of those four variables.

Briefly justify your choice of method to select the group. Be sure to include a justification for any options that you had to specify in carrying out the analysis and give any output you used to make your decision.
4) Consider the two subsets of variables:

“Body measurements” = Length, ChestG, and Weight

and

“Head measurements” = HeadL, HeadW, and NeckG

Perform the appropriate multivariate analysis to determine what the strongest linear relationships are between these two sets of variables.

Briefly describe what those relationships are in terms of the variables, say how you measured if the relationships were statistically significant, give a measure of the strength of the relationships, and say if you think the relationships are very strong or not.

5) One way of graphically displaying the relationships between the bears would be to construct a “map” of them based on the ten quantitative variables related to size (Weight to sNeckG).

Determine the “best” number of dimensions for capturing and displaying the results using isometric multidimensional scaling with Karl Pearson distance. There are likely several possible number of dimensions that could be reasonable, justify your choice.

For the two-dimensional scaling produce a plot of the bears labeled by Sex, and a plot labeled by Maturity. Does either of these two variables seem to be separated very well in the map?

6) It is desired to see if the four sex-maturity groupings of bears differ on average in terms of the five variables sChestG, sHeadL, sHeadW, sNeckG, and Length.

State the appropriate null and alternate hypothesis for these vectors of measurements and conduct the hypothesis test. Briefly say why you chose the test statistic you did. Give the output and state your conclusion.

Why should you worry about violations of the equal covariance assumption for this data set? (You do not need to check the assumptions).

Graduate Students, also answer one of the following:

2b) There is a “seemingly unrelated” method that can be used to answer questions similar to both MANOVA and discriminant analysis. Name the method for the two group case, give the general equation that would describe the relationship between the two groups and the variables (being sure to identify any symbols you use in the equation).

5b) In problem 5 the plot of the first two dimensions may be somewhat different for the 2 and 3 dimensional solutions. Briefly explain why this cannot happen for classical multidimensional scaling.