STAT 530 - Fall 2003 - Homework 6

Due: Monday, November 24th

1) Problem one uses the data set:
`http://www.stat.sc.edu/~habing/courses/data/crabs2.txt`
that is related to the data set `crabs` we used in class.

The data set contains four groups of fifteen crabs each: `O` = orange male, `o` = orange female, `B`= blue male, `b` = blue female.

Each crab has eight measurements. The first four are: `FL` = frontal lobe size (mm), `RW` = rear width (mm), `CL` = carapace length (mm), and `CW` = carapace width (mm). The next four variables are indicated with the same names as the previous four, but with an `s` added at the beginning. These were created by first dividing each of the original measurements by the total body depth (a measure of overall size). They were then standardized by subtracting the mean and dividing by the standard deviation.

a) If you want to tell different species of crab apart using cluster analysis, why might you want to divide the four individual measurements by the overall size?

b) If you want to tell different species of crab apart using cluster analysis, why might you want to standardize the measurements?

c) Conduct a cluster analysis using the four `s` variables, indicting each crab on the dendogram by its group, using each of nearest neighbor, average, and furthest neighbor linkages. Which seems to do the best job of separating the four groups? (Briefly say why, maybe by labeling the picture.)

d) Compare this to using the linkage you found best and the non-`s` variables. Did the `s`-variables indeed seem to work better? (Briefly say why, maybe by labeling the picture.)

2) The second problem uses the data set
`http://www.stat.sc.edu/~habing/courses/data/scdatamod.txt`
with South Carolinians extracted from
`http://www.cdc.gov/brfss/technical_infodata/surveydata/2002.htm#data`

While linear discriminant analysis did not enable us to tell different racial groups apart using the foods people ate, perhaps canonical correlation analysis can find for us relationships between peoples diets and there physical characteristics.

Conduct a canonical correlation analysis using `AGE` (in years), `HEIGHT` (in inches), `WEIGHT` (in pounds), and `BMI` for one set of variables and the servings per week of `FRUITJUICE`, `FRUIT`, `GREENSAL`, `POTATOES`, `CARROTS`, and `VEGETABLES` as the other set.

a) Report how many pairs of canonical covariates were statistically significant at α=0.05.

b) Give brief name/descriptions to each of the significant canonical covariates.

c) Report the r^2 for each of the significant pairs. What about this data set makes it that such small correlations can be statistically significant?