Homework 7 Statistics 8320 (Wikle) April 28, 2016

DUE: May 5, 2016

1. Consider the following correlation matrix that is based on data concerned with the development of a standardized scale to measure beliefs about controlling pain. A sample of 123 people suffering from extreme pain were asked to rate nine statements about pain on a scale from 1 to 6, ranging from disagreement to agreement with the statement. The nine statements were:

   - (1) Whether or not I am in pain in the future depends on the skills of the doctors.
   - (2) Whenever I am in pain, it is usually because of something I have done or not done.
   - (3) Whether or not I am in pain depends on what the doctors do for me.
   - (4) I cannot get any help for my pain unless I go to seek medical advice.
   - (5) When I am in pain, I know that it is because I have not been taking proper exercise or eating the right food.
   - (6) People's pain results from their own carelessness.
   - (7) I am directly responsible for my pain.
   - (8) Relief from pain is chiefly controlled by the doctors.
   - (9) People who are never in pain are just plain lucky.

   The correlation matrix for these responses is:

1.0000								
-.0385	1.0000							
.6066	-.0693	1.0000						
.4507	-.1167	.5916	1.0000					
.0320	.4881	.0317	-.0802	1.0000				
-.2877	.4271	-.1336	-.2073	.4731	1.0000			
-.2974	.3045	-.2404	-.1850	.4138	.6346	1.0000		
.4526	-.3090	.5886	.6286	-.1397	-.1329	-.2599	1.0000	
.2952	-.1704	.3165	.3680	-.2367	-.1541	-.2893	.4047	1.0000

   Use PROC FACTOR (reading in the correlation matrix directly) to perform a factor analysis on the pain data. To read the correlation matrix, given in the file: `pain.dat` do the following:

```
data pain (type = corr);
infile 'pain.dat' missover;
input _type_ $ _name_ $ p1 - p9;
run;
```

(a) Use the MLE method, extracting two factors. Are two factors sufficient? Justify your answer.

(b) Redo the MLE factor analysis with 3 factors and a varimax rotation. Are three factors enough? Interpret the rotated factors.

(c) Use the principle factor method with varimax rotation. How does the interpretation of the first 3 factors differ from the MLE/varimax analysis?

(d) Use the principle factor method and an oblique rotation (PROMAX); compare results.

2. This problem concerns the detection of SPAM emails. In particular, the file `spamdetect_train.dat` contains 4101 records containing a response of 0 (not spam) or 1 (spam) in the last column, and 57 attributes (covariates) described in the file `spam_names.txt`. Your task is to perform discriminant analysis that can classify spam based on these attributes. I also include a test dataset in the file `spamdetec_test.dat`.

   (a) Are the covariance matrices the same for spam and not spam? What does this imply about the type of discriminant analysis that should be performed?

   (b) Perform a linear discriminant analysis. Report the cross-validiation summary and the results from the test dataset.

   (c) Perform a quadratic discriminant analysis. Report the cross-validation summary and the results form the test dataset.

   (d) Use the PROC STEPDISC procedure to find a good reduced-variable discriminant model. Report which variables are included and why you chose the number that you did. How do the probabilities of classification (on the test sample) compare to the models above?

3. Consider the data on air pollution in some US cities given in `usair.dat`. The columns correspond to: (1) SO_2 content of air, (2) average annual temperature, (3) number of manufacturing enterprises employing 20 or more workers, (4) population size (1970 census), (5) average annual wind speed, (6) average annual precipitation, (7) average number of days per year with precipitation.

   (a) Consider a hierarchical cluster analysis of variables 2-7 using “complete” linkage (use standardized data). Give a plot of the dendrogram and discuss how many clusters you think would be appropriate to use with this analysis (or, if it isn’t clear, give an indication as to why).

   (b) Perform a K-means cluster analysis using variables 2-7 and assume 4 clusters. Using the cluster categories as a class variable, use proc glm in a 1-way ANOVA to see if there is a significant difference among the SO_2 variable given the clusters. Interpret your results. Note, in the following SAS code, the output file “clustout” contains the cluster categories in the “cluster” variable.

```
proc glm data=clustout;
   class cluster;
   model so2=cluster;
 run;
```