University of the Philippines
Mathematic Department

2004-2005
Master

Data learning 6 : Introduction to SAS software (continued)

Exercice 1 : Example of Cluster Analysis (HA). With the help of the notes of the lecture 2, perform a cluster analysis of the data set "Cities.txt" that shows the flying Mileages between 10 American cities : first import the data file in SAS, then perform a HA using the procedure `CLUSTER`. Plot the dendrogram. How many clusters would you keep ?

Exercice 2 : Example of Cluster Analysis (K-means). The file called "Fischer.sas" contains the SAS code of a cluster analysis using K-means method with $q = 2$ (et $q = 3$) of a welknown data set studied by Fischer in 1936. This data set consists in the measures in mm of sepal length, sepal width, petal length, and petal width on 50 iris specimens coming from three different species called setosa, versicolor and Virginica. The file contains also instructions needed for create a plot of the cloud, in the case of $q = 3$, in the plan of the two first canonical components, using three different colors for the three different clusters.

Open this file in SAS and run it. Try to interprete the different outputs and, reading the code, to understand how each of these outputs have been produced.

Exercice 3 : Simulation of a cloud with two clusters Read the file called "simulation.sas" : it create a data set called "cloud" consisting in 10 points randomly distributed around the point $(-1, 2)$ and 20 points randomly distributed around the point $(2, -1)$ (Notice that `rannor(12345)` drows a number according to the $\mathcal{N}(0, 1)$ distribution) ; then, using a cluster analysis, the code try to recover the partition of the set between the 10 and the 20 points. Run the code a first time.

1. Change the centers $(-1, 2)$ and $(2, -1)$ and see what happens : try cases where the two clusters are well separated and cases where they are not.
2. Change the number of points in the two clusters.
3. Create more than two clusters
4. Try other "methods" than Ward and compare.

Exercice 4 : Regression using Analyst. It is very easy to do any regression in SAS using "Analyst". Recall that to enter this program, select `Solution/Analysis/Analyst`. To create a sample data in the sasuser directory, choose `Tools/Sample Data` and select the file called "Fitness". Then select `File/Open by SAS Name` and then `Fitness/OK`.

The data set consists in 8 variables measured on several men taking a physical fitness course and belonging to various groups. The variables are age, weight, oxygen (oxygen intake rate per kg per mn), runtime (nb of mn taken to run 1.5 miles), rstpulse (heart rate while resting), runpulse (heart rate while running), maxpulse (maximum heart rate while running), group.

Have a look on the data set (than close the Viewtable). To request a linear regression analysis, select `Statistics/regression/Linear` and run several times, changing the outputs as you ask for. Is there a linear model (and which one ?) that fit well the data ?

Exercice 5 : Logistic regression using Analyst. Same exercice as the previous one with the file called "Coronary2" that you can create as a sample data. It contains a variable called ca (for desease Yes/No) that has to be explained by three explanatory variables, sex, ecg (ST segment depression) and age.

Exercice 6 : Regression without Analyst. Without using Analyst (but using the procedure REG directly), find the best linear model for the data set called "wtht.txt" in order to predict weight from hight and age. Are you sure that the model fits well the data ? Drow a plot of your data together with their linear model either in the weight/hight plan or in the weight/age plan.