STA 402/502 Homework 4

Due: September 28th (Friday), before class

Please read the homework guidelines before working on the homework. Homework that does not follow the guidelines will be deducted points from now on. You are to complete this assignment on your own. Remember to include an intro comment block on all programs written. Each problem should be attempted as its own program.

1. The dataset "data_set_ALL_AML_independent.csv" from canvas website contains the gene expression measurements corresponding to ALL and AML samples from Bone Marrow and Peripheral Blood (Golub et al., 1999). The rows correspond to different genes and the columns correspond to different patients.

   (a) Use appropriate method to read the dataset into SAS as a SAS dataset. Double check to make sure you have done this correctly.

   (b) Write a single DATA step that takes an existing SAS data set and creates a new SAS data set for which each observation consists of the mean, median and standard deviation of the absolute values of all numeric variables in the corresponding observation from the original set. (Note: the new data set should have the same number of observations as the old data set, but only three variables.) Your code should employ macro variables so that **it works for any SAS data set**.

   Hint: Use arrays with the functions MEAN, MEDIAN, STD, and ABS for the numerical calculations.

   (c) Apply your code to the data set you read in from part (a), use PROC PRINT to show the 3567-3570's observations of the resulting dataset.

2. The official results of the 2012 London Olympics mens 3-meter springboard diving finals can be found in the SAS data set called "diving.sas7bdat". The data consist of six observations per diver, one for

each of their six dives in the final event. The variables in this file are diver's name, country, height (m), weight (kg), dive number (1 to 6), dive code, degree of difficulty, description, position, scores from each of seven judges and penalty.

Before working on the exercises below, examine this SAS data set including the variable name, labels and other attributes. (Variable attributes can be found by right click on the variable name.)

(a) Calculate the diving score for each of the observations, create a new variable containing this calculated score. Follow the calculation method below:

From the scores provided by the seven judges, cross out the two highest and the two lowest scores. Then add the rest of the scores together.

Each attempted dive has a degree of difficulty calculated in advance. This is based on many factors, such as the number of twists and somersaults and the take-off and entry positions. Multiply your last sum by the degree of difficulty to get the final score for this dive.

Hint: 1. If you open a SAS dataset and look at the column names, they're actually the "label" of the variables. You can not use the label directly in your program to refer to SAS variables. To find the SAS variable name, look at the "name" in column attributes. 2. SAS functions SMALLEST and LARGEST maybe useful for your calculation.

(b) Follow the requirements below, create a .txt file containing all the observations from the updated SAS dataset. Open the .txt file using a simple editor like notpad. Include a screen-shot of the first 18 observations from this dataset in your homework.

Requirements:

- Only include the name, country, height, weight, dive number and the calculated score in the .txt file. You don't need to include the variable name as the first row in the .txt file.
- Make sure the observations for each of the variables are well separated. For each variable, all the observations should start at the same column.

- For variables height, weight and the calculated score, print 2 decimal points for each observations. You may search "PUT Statement, Column" in the SAS support page for syntax and example.

Reference

T.R. Golub, D.K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J.P. Mesirov, H. Coller, M. Loh, J.R. Downing, M.A. Caligiuri, C.D. Bloomfield, and E.S. Lander (1999), Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression, *Science*, 286:531-537.