Homework 1: Introduction to Statistical Computing Using SAS

Instructor: H. Peng
STAT 42100: Modern Statistical Modeling Using R and SAS
STAT 52100: Statistical Computing

Instruction: Collaboration on homework assignments is acceptable, but all write-ups must be done independently and clearly indicate the submitter's understanding of the material. Unclear or disorganized homework may have points removed, even if the content is correct. **No hardcopy of the typed report needs to be turned in. The pdf file should be submitted electronically on canvas.** An assignment handed in after the deadline is late, and may or may not be accepted. My solutions to the assignment questions will be available when everyone has handed in their assignment.

Edit the program(s) and output together into a single document, showing the lines of code and relevant output produced by SAS/R. Your answers must be easy for the grader to find. A simple structure is, for each part of each question in order, to put these three things:

- Your code
- Your output
- Your answers and explanation

For SAS, the code and output are naturally separate, so this order is good; for R, the code and output can be intermingled, and that is OK. If your assignment is disorganized or otherwise difficult for the grader to deal with, you can expect to lose marks.

You are reminded that work handed in with your name on it must be entirely your own work. It is as if you have signed your name under it. If it was done wholly or partly by someone else, you have committed an academic offence, and you can expect to be asked to explain yourself. The same applies if you allow someone else to copy your work. The graders will be watching out for assignments that look suspiciously similar to each other (or to my solutions). Besides which, if you do not do your own assignments, you will do badly on the exams.

Problem 1 The problem uses US Medicare payment data, available on canvas. The data are documented in this file. You will need to read the data documentation closely to fully understand this, but briefly, each line in the data file represents a number of similar services provided to various patients by a single "provider" (usually a doctor, nurse, pharmacist, etc.). Only the mean and standard deviation of the payment amounts is given, along with some information about the provider, and about the type of service.

Note: You will need to carefully read the data documentation to understand how the questions given below relate to the data in the file.

In this exercise you will perform some basic manipulations of this dataset in SAS.

a) Write a SAS program using a data step to load the data into SAS. You will need to search the web to see how to set the delimiter for a tab-delimited file. Print the first 10 rows and run proc contents to verify that the data are loaded correctly. Note that you will need to use the dsd option on the infile line to correctly handle empty fields, and you must set the proper delimiter for tab-delimited files. Note just load the following variables (the variable followed by a dollar sign is a character variable, otherwise numerical)

```
NPI NPPES_CREDENTIALS $ NPPES_PROVIDER_GENDER $ NPPES_ENTITY_CODE $
NPPES_PROVIDER_ZIP $ NPPES_PROVIDER_STATE $ PROVIDER_TYPE $
MEDICARE_PARTICIPATION_INDICATOR $ PLACE_OF_SERVICE $ HCPCS_CODE $
HCPCS_DRUG_INDICATOR $ LINE_SRVC_CNT BENE_UNIQUE_CNT BENE_DAY_SRVC_CNT
AVERAGE_MEDICARE_ALLOWED_AMT STDEV_MEDICARE_ALLOWED_AMT
AVERAGE_SUBMITTED_CHRG_AMT STDEV_SUBMITTED_CHRG_AMT AVERAGE_MEDICARE_PAYMENT_AMT
STDEV_MEDICARE_PAYMENT_AMT;
```

b) Use proc freq to determine the frequency of claims within each state that were paid to female and male providers. Which states had the least and the greatest proportion of claims paid to females?

c) Reduce the dataset to have one record for each provider (hint: you can do this via a data step with the first.xxx syntax). Then repeat part (b) using this dataset.

Problem 2 Download the Excel files for the six groups from canvas. You should have the following variables in each Excel file:

Variable Name	Description	Type
GROUP	Group Number	Numeric
ID	ID Number within Group	Numeric
RAN	1= Ran, 0=Did not run	Numeric
AGEYR	Age in years	Numeric
AGEMO	Age in months	Numeric
SEX	M or F	Character
HR1	Heartrate at time 1	Numeric
HR2	Heartrate at time 2	Numeric

1. Read all 6 groups of Excel data into SAS.
2. Get descriptive statistics for the numeric variables in all six of the datasets using Proc Means.
3. Combine the data sets for all groups by stacking them into one temporary data set, using a set statement in SAS. Get Proc Contents for the Allgroups data set, with the variables in creation order, by using the varnum option.
4. Get descriptive statistics for all numeric variables in your Allgroups dataset using Proc Means.
5. Use a class statement to get descriptive statistics for each GROUP.
6. Get frequencies for the categorical variables GROUP, SEX, and RAN, using Proc Freq.
7. Save your dataset as a permanent SAS dataset, using commands similar to those shown below:

```
libname b510 "/home/hlwang1/";
data b510.allgroups;
   set work.allgroups;
run;
```