STAT 440 – Homework 7

Students are encouraged to work together on homework. However, sharing or copying any part of the homework is an infraction of the University's rules on Academic Integrity.

Final submissions must be uploaded to our Compass 2g site on the Homework page. No email, hardcopy, or late submissions will be accepted.

Getting the program file ready

a. Create a folder on the hard drive with the following pathname – C:\440\hw7. Save all data files accompanying this assignment in that folder. If you cannot create the folder because you are working on a university computer and don't have permission, create the …\440\hw7 folder elsewhere.
b. Assign the library reference **hw7** to the folder 'C:\440\hw7'. Use this library as your permanent library for this assignment. If you could not create the folder, assign the library reference **hw7** to your …\440\hw7 folder.
   Note: If you are using a folder other than 'C:\440\hw7', you must change any pathname references in your program file to 'C:\440\hw7' before submitting your homework.

Submitting your work to Compass 2g

You are to submit two (and only two) files for your homework submission.

1. Your SAS program file which should be saved as **HW*n*_*YourNetID*.sas**. For example, my file for the HW7 assignment would be HW7_dunger.sas. All program statements and code should be included in one program file.

2. Your Report including all relevant output to address the exercises. For this homework, use ODS to send your results to a Rich Text Format (RTF) file called ***YourNetID*_HW*n*.rtf**. Only include your final set of output. Do not include output for every execution of your SAS program. Use the template file **hw5 template.sas** as your guide.

You have an unlimited number of submissions, but only the last one will be viewed and graded. Homework submissions must always come as a pair of files, as described above.

1. The Consumer Expenditure Survey (CE) is conducted by the Bureau of Labor Statistics to provide data on the buying habits of American consumers. The Interview data you'll explore generally tracks consumer units' (CU) large expenditures, such as major appliances and cars. An Interview "quarter" refers to the calendar quarter in which the interview occurred. For example, any Consumer Unit interviewed in April, May, or June would have their data stored in the quarter 2 (2014Q2) datasets. During an interview, the CU is asked to report expenditures for a reference period of three months. So, for a CU interviewed in April, their expenditures in the YYQ2 file are for January, February, and March.

   The Interview survey collects data at each quarter of the year at both the consumer unit (i.e., family) level and member (i.e., individual person) level. Thus, each consumer unit (CU) may be composed of multiple members (i.e., a family could have 1, 2, 3,… members). A CU may or may not participate in all the interviews (e.g., respond to 1st and 4th quarters, but skip 2nd and 3rd).

   You will use the following SAS data sets.
   fmli14*i*
   - There is one record per CU.
   - Each CU is uniquely identified by NEWID.
   - It is possible for a CU to skip an interview. For example, a CU could have a 2nd, 3rd and 5th interview but no 4th interview.
   - Variables include demographics for the reference person and spouse of reference person, income at the CU level, sample housing unit information, and summary level expenditures.

   memi14*i*
   - There are multiple records per CU.
   - There is one record per member.
   - Unique records are defined by the combination of NEWID and MEMBNO.
   - Variables include demographics about CU members, member level income, and member relationship status to the reference person.

   Description:
   - The specifications of each variable in each data file can be found in the **Interview Dictionary** file. It contains information on every one of the hundreds of variables from the original survey, but only a subset of those variables are used in the data sets provided.
   - In the **fmli** data sets, NEWID is unique to each observation. That is, a valid NEWID occurs at most once in each of the four **fmli** data sets.
   - In the **memi** data sets, NEWID may occur more than once if the CU (i.e. household) has more than one member. For example, a family of four would share the same NEWID and so those four observations in a **memi** data set would all have the same NEWID.

   a. Use PROC CONTENTS to view the descriptor portion of each of the eight data sets. Construct a table that lists the number of observations and number of variables in each data set. This table can be made in Word and does not have to be compiled using SAS. (Include the table in the HW Report. Do not include the output from PROC CONTENTS.)
      - Note that all the **fmli14** data sets have the same number of variables, as do the four **memi14** data sets.

b. Look at the Variable Attributes table for one of the **fmli** and one of the **memi** data sets. Using the **Interview Dictionary** file, create a user-defined format for each variable that would benefit from having a format to interpret its levels.
    - Note that you might not need a unique format for each variable. A format can be applied to multiple variables.
    - You should be able to do some copying-and-pasting from the **Interview Dictionary** file to save time on typing.

c. Concatenate (but do not interleave) the four family-level data sets. Also create a new variable called QTR that uniquely identifies during which quarter of 2014 the interview took place. Name the resulting temporary data set **fmli2014*_NetID***.
    - Apply SAS and user-defined formats as needed.

d. Print the descriptor portion of the new data set. (Include your results in the HW Report.)
    - Use the results of part (a) to check the math of your concatenation.

e. Create a two-way cross-tabulation table of EDUC_REF (rows) and SEX_REF (columns). (Include results in the HW Report.)
    - Suppress the row and column percentages.

f. Print a table containing the median, mean, and standard deviation for CU income before taxes (FINCBTAX) as partitioned by REF_RACE. (Include results in the HW Report.)

g. Concatenate (but do not interleave) the four member-level data sets. Also create a new variable called QTR that uniquely identifies during which quarter of 2014 the interview took place. Name the resulting temporary data set **memi2014*_NetID***.
    - Apply SAS and user-defined formats as needed.

h. Print the descriptor portion of the new data set. (Include your results in the HW Report.)
    - Use the results of part (a) to check the math of your concatenation.

i. Merge the data sets **fmli2014*_NetID*** and **memi2014*_NetID*** into a new permanent data set called **ce2014*_NetID***. This should be a merge that matches a consumer unit with all corresponding family member interview responses.

j. Print the descriptor portion of the new data set. (Include your results in the HW Report.)
    - Use the results of part (a) to check the math of your concatenation.

k. Ideally, the number members/observations in each family (CU) in the **memi** data should match the FAM_SIZE variable in the **fmli** data for the same interview quarter. Check this provision by validating the data.
    - No need to clean the data if you find inaccuracies.
    - Hint: Keep in mind that you can output PROC FREQ results to a data set.