Homework 2: Compiling Datasets

EPID 640 Fall 2017

This homework assignment begins our replication of **Bansil et al., 2011** using the 2013-2014 data from NHANES. You may discuss the assignment with others but you MUST run your own code and write up the answers in your own words. Please **turn in a printed copy** of your write-up and SAS commands (attached as a printout to the back of your write-up) in class at **9:30AM on Friday, September 29th** and **submit an electronic copy** of your write-up and code via the CANVAS Assignment tab. ONLY output that is **relevant to the questions** should be included.

Be sure that: 1) Your SAS code runs from start to finish,

2) Your results make sense (check your sample sizes and look for unreasonable, unlikely, or impossible answers),

3) Your homework write-up has your name on EACH page. Submit your write-up on Canvas as an attachment to Assignment 2. Your file should be turned in to Canvas as a .doc, .docx, or .pdf in the following format: LASTNAME\_FIRSTNAME\_HW2.docx

4) Your code is well commented. Commenting is done by **\***type in any text here**;** or **/\*** type in text here **\*/**. Be sure to include the homework number and your name at the top of your homework code, identify each question in the code, and describe each new task. Also, format your code (indentation and carriage returns) to improve readability. Your code file should be turned in to Canvas as a .sas file in the following format: LASTNAME\_FIRSTNAME\_HW2.sas

5) 5% will be deducted if either of tasks 3 or 4 above is not completed.

6) All files needed for the homework can be found in the Files section of Canvas in the folder Homework 2 (including the homework questions, NHANES documentation files [.htm files], and datasets).

**CREATING A CODE BOOK FOR THE SEMESTER**

1. For this question we will create a codebook that will be used throughout the rest of the semester in future homework assignments. Use the files listed below to expand this table to include **a complete list of all the variables that will be used in our analysis.** Include all variables mentioned in the paper including those needed to check the inclusion/exclusion criteria as well as their units and possible values. The following codebooks have been included with the homework to help you identify the variables needed:

DEMO\_H (Demographics)

BPX\_H (Blood Pressure Measures)

SLQ\_H (Sleep Disorders)

BPQ\_H (Blood Pressure Medication)

BMX\_H (Body Measures)

DIQ\_H (Diabetes)

SMQ\_H (Smoking)

HIQ\_H (Health Insurance)

These can also be found on the NHANES website:

https://wwwn.cdc.gov/nchs/nhanes/continuousnhanes/default.aspx?BeginYear=2013

|  |  |  |  |
| --- | --- | --- | --- |
| Variable | Variable Name | Possible Values (Defined) | Dataset |
| ID Number | SEQN |  | all datasets |
| Sex | RIAGENDR | 1: Male  2: Female  .: Missing | DEMO\_H |
| Age | RIDAGEYR | 0 to 79: Range of values  80: 80 years of age and older  .: Missing | DEMO\_H |
| Race | RIDRETH1 | 1: Mexican American  2: Other Hispanic  3: Non-Hispanic White  4: Non-Hispanic Black  5: Other Race - Including Multi-Racial  .: Missing | DEMO\_H |
| Education (for adults) | DMDEDUC2 | 1: Less than 9th grade  2: 9-11th grade (includes 12th grade with no diploma)  3: High school graduate/GED or equivalent  4: Some college or AA degree  5: College graduate or above  7: Refused  9: Don’t know  .: Missing | DEMO\_H |
| Poverty to Income Ratio | INDFMPIR | 0-4.99: Range of values  5: Value greater than or equal to 5.00  .: Missing | DEMO\_H |
| Pregnancy Status | RIDEXPRG | 1: Yes, positive lab pregnancy test or self-reported pregnant at exam  2: The participant was not pregnant at exam  3: Cannot ascertain if the participant is pregnant at exam  .: Missing | DEMO\_H |
| First Systolic Reading | BPXSY1 | 66-228: Range of values  .: Missing | BPX\_H |
| Second Systolic Reading | BPXSY2 | 66-230: Range of values  .: Missing | BPX\_H |
| Third Systolic Reading | BPXSY3 | 62-228: Range of values  .: Missing | BPX\_H |
| Fourth Systolic Reading | BPXSY4 | 80-212: Range of values  .: Missing | BPX\_H |
| First Diastolic Reading | BPXDI1 | 0-122: Range of values  .: Missing | BPX\_H |
| Second Diastolic Reading | BPXDI2 | 0-116: Range of values  .: Missing | BPX\_H |
| Third Diastolic Reading | BPXDI3 | 0-118: Range of values  .: Missing | BPX\_H |
| Fourth Diastolic Reading | BPXDI4 | 0-128: Range of values  .: Missing | BPX\_H |
| Sleep Disorder | SLQ060 | 1: Yes  2: No  7: Refused  9: Don’t know  .: Missing | SLQ\_H |
| Sleep Duration | SLD010H | 2-11: Range of values  12: 12 or more hours  77: Refused  99: Don’t know  .: Missing | SLQ\_H |
| Sleep Quality | NA for 2013-2014 |  |  |
| Self-Report Hypertension | BPX020 | 1: Yes  2: No  7: Refused  9: Don’t know  .: Missing | BPQ\_H |
| Told by doc to take drug | BPX040A | 1: Yes  2: No  7: Refused  9: Don’t know  .: Missing | BPQ\_H |
| Currently taking drug | BPQ050A | 1: Yes  2: No  7: Refused  9: Don’t know  .: Missing | BPQ\_H |
| Body Mass Index (kg/m2) | BMXBMI | BMI range: 12.1-82.9  .: Missing | BMX\_H |
| Diabetes | DIQ010 | 1: Yes  2: No  3: Borderline  7: Refused  9: Don’t know  .: Missing | DIQ\_H |
| Smoked 100 Cigarettes | SMQ020 | 1: Yes  2: No  7: Refused  9: Don’t know  .: Missing | SMQ\_H |
| Current Smoker | SMQ040 | 1: Every day  2: Some days  3: Not at all  7: Refused  9: Don’t know  .: Missing | SMQ\_H |
| Health Insurance | HIQ011 | 1: Yes  2: No  7: Refused  9: Don’t know  .: Missing | HIQ\_H |

**READING IN AND SAVING DATASETS**

2. Create a permanent SAS library called **nhanes**.

1. Paste a copy of the log output from the LIBNAME statement into your homework.

LIBNAME nhanes "C:\Users\smecham\Desktop\nhanes";

RUN;

b) Describe the purpose of creating a SAS library.

To be able to re-access your work. If you don’t create a SAS library, files will be saved to the temporary work folder and will disappear after you end your session and you won’t be able to open up your previous files.

1. Import the demographic data (**DEMO\_H.xlsx**) using a PROC IMPORT statement. You can do this using the import wizard in SAS to generate the code or coding the program directly. Include a copy of the PROC IMPORT statement in your SAS code program. Save as a dataset called **demo** in your **nhanes** library. Review your log file and open the dataset to be sure that all imported properly.

1. How many observations and variables are in the file? (Hint: You can right click on the file and request properties if your log does not display the information.)

10,175 observations and 47 variables

1. In the PROC IMPORT statement, what is the DBMS option used?

XLSX

1. According to the code book, how would the age for a person who was 95 be coded (i.e. what the age would appear to be in the demo dataset if a person is 95 years old)?

80

1. Import the dataset that contains sleep disorders (**SLQ\_H.txt**) using a DATA step. (A trick of the trade is to open the file, copy and paste the header row so as not to have to type in the variable names. You can do this on the remaining questions too.) First import the file into a dataset called **sleep** in your temporary work space. Once you are content that the dataset imported properly (check your log and dataset), re-write and run the code to save it into the **nhanes** library. Your code should include BOTH your temporary and permanent importing steps.
2. What was the delimiter of this file?

Tab

1. How many observations and variables are in the file?

6461 observations and 4 variables.

1. Import the smoking data (**SMQ\_H.csv**) using a DATA step. As before, first import the file into a dataset called **smq** in your temporary work space. Once you are content that the dataset imported properly (check your log and dataset, hint: remember that not all variables are numeric), re-write and run the code to save it into the **nhanes** library. Your code should include BOTH your temporary and permanent importing steps.
2. What options, if any, are required in the INFILE statement for this dataset?

Delimiter option as comma delimited, Firstobs option to take into account header row, and the DSD to take into account delimiters with missing data.

1. How many observations and variables are in the file?

7168 observations and 32 variables

1. Import the blood pressure medication data (**BPQ\_H.xpt**) as a SAS Export File. These are nice ways to package SAS files for easy reading in any version of SAS. This is the true format of the NHANES data. To read a SAS Export file you must type the following commands:

LIBNAME BPQ XPORT "***yourfilepath***\BPQ\_H.xpt";

**PROC** **COPY** IN=BPQ OUT=nhanes;

**RUN**;

*!!!! NOTE – occasionally there are incongruities between Microsoft Word and SAS and the quotation marks are not read correctly. Should you experience a bizarre error try deleting the quotations and manually typing the quotations back in.*

Be sure to replace **yourfilepath** in the XPORT command with your local directory name where the export file is stored. Here the LIBNAME statement tells SAS the directory of the xport file and the PROC COPY statement tells SAS to get the data from the IN directory (the directory of the xport file) and put it in your OUT directory (nhanes for us). Look at your log and dataset file.

1. How many observations and variables are in the file?

6464 observations and 14 variables

1. Look at the SAS dataset. Describe how the variable names (column headers) appear in the table?

They appear as full descriptions, not a code number. For example, instead of ‘SEQN’ for patient sequence number, in this table it appears ‘Respondent sequence number’.

1. Right click on the dataset in your explorer window and select the view columns option. Scroll through the window. Can you figure out why the display of the variables appears to violate some basic rules of variable names?

When processing the data, SAS used the typical coded variable names, but labels with longer descriptions were inputted so that SAS would automatically replace the variable name shortcuts with a full description in the output.

1. Open the dataset. Why is there so much missing data (.) in this dataset? We have included the data dictionary with the homework (**BPQ\_H**).

There is so much missing data because anyone that reported “2” to the first question, representing that they have never been told they had high blood pressure, didn’t answer the rest of the questions about blood pressure. Therefore any of the other columns in the people that marked “2” at first will be marked as missing in SAS.

1. Of the datasets you have imported into SAS, which file has the fewest number of observations in it and which has the most? Recalling that these are from the same survey, what might be an explanation for different numbers of observations in each file? (Hint: You may want to review the documentation files for clues.)

The sleep disorders dataset had the fewest numbers of observations and the demographic dataset had the most. The files that have fewer observations are related to the amount of questions involved in each section. The demographic dataset included a lot of variables like age, sex, race, education level, etc. whereas the sleep disorder survey only asked about whether or not the individual had a sleep disorder and how long they typically sleep. Because there are less questions asked in the sleep disorder portion of the survey, it makes sense that that dataset will have much fewer observations than a more in depth portion like demographics.

1. I’m sure that you’ve noticed that there are many missing data, refused data, and unknown data in the NHANES datasets.
2. What coding has been used to identify refused and don’t know answers for how often checked blood pressure at home in the last 12 months (blood pressure medication dataset)?

7777: Refused

9999: Don’t know

1. Why do you think that values like these have been chosen instead of a character string like “Refused” or a refused code value of 7, for example?

The range of values one could report is from 1 through 2704, so if you used a value of 7, there would be no way of knowing if that person reported taking their blood pressure at home 7 times in the past 12 months or if they are refusing to answer the question. Refused is not a numerical phrase, and you need the responses to be either all numeric or all characters for SAS to run the analysis correctly.

1. What impact would these values have on our analysis if we neglected to account for them in our code (i.e. if we completed our analysis leaving them unchanged)?

If we used a code of 7, our analysis would be inaccurate because we would be analyzing all the people who refused to answer that question as if they had taken their blood pressure 7 times in the past year. It would also make it impossible to determine how many people refused to answer that question. If you use a character string for a numeric variable, SAS will try to convert the characters to numbers and you will receive an error, so SAS won’t be able to analyze the dataset correctly.

1. There is no single variable for smoking status in the smq\_h dataset that is coded as current, former, and never. What two variables in the smoking dataset should we use to determine if participants are current, former, and never smokers (hint: review the codebook for smq\_h included with the homework)? Describe your thinking about how we might combine the information from these two variables into one variable.

The two variables we need are the SMQ040 (do you now smoke cigarettes) and SMD030 (age started smoking cigarettes regularly). To determine current smokers, add up the respondents that reported ‘every day’ and ‘some days’ for the “do you now smoke cigarettes’ question. You can determine the former smokers by taking the respondents that reported a particular age value for the “age started smoking cigarettes regularly” question. You can also use that variable to determine the never smokers, because the never-smokers would have reported “never smoked cigarettes regularly” on that question.

[CODE ON NEXT PAGE]

Code:

\*Homework 2: Compiling Data Sets - Stephanie Mecham | EPID 640 Section 4;

\*Question 2 - Creating a permanent library;

LIBNAME nhanes "C:\Users\smecham\Desktop\nhanes";

RUN;

\*Question 3 - Importing demographic data using PROC IMPORT;

proc import

datafile='C:\Users\smecham\Desktop\nhanes\DEMO\_H.xlsx'

out= nhanes.demo

dbms= xlsx

replace;

getnames=yes;

datarow=2;

run;

\*Question 4 - Import sleep disorder data using DATA step;

\*saving code to temporary work space;

data sleep;

infile 'C:\Users\smecham\Desktop\nhanes\SLQ\_H.txt'

DLM='09'X

firstobs= 2;

input SEQN SLD010H SLQ050 SLQ060;

RUN;

\*saving code to permanent work space;

data nhanes.sleep;

infile 'C:\Users\smecham\Desktop\nhanes\SLQ\_H.txt'

DLM='09'X

firstobs= 2;

input SEQN SLD010H SLQ050 SLQ060;

RUN;

\*Question 5 - Importing smoking data using a DATA step;

\*saving code to temporary workspace;

data smq;

infile 'C:\Users\smecham\Desktop\nhanes\SMQ\_H.csv'

DLM=','

DSD

firstobs=2;

input SEQN SMQ020 SMD030 SMQ040 SMQ050Q SMQ050U SMD055 SMD057 SMQ078 SMD641 SMD650 SMD093 SMDUPCA $ SMD100BR $ SMD100FL SMD100MN SMD100LN SMD100TR SMD100NI SMD100CO SMQ621 SMD630 SMQ661 SMQ665A SMQ665B SMQ665C SMQ665D SMQ670 SMQ848 SMQ852Q SMQ852U SMAQUEX2;

run;

\*saving code to permanent workspace;

data nhanes.smq;

infile 'C:\Users\smecham\Desktop\nhanes\SMQ\_H.csv'

DLM=','

DSD

firstobs=2;

input SEQN SMQ020 SMD030 SMQ040 SMQ050Q SMQ050U SMD055 SMD057 SMQ078 SMD641 SMD650 SMD093 SMDUPCA $ SMD100BR $ SMD100FL SMD100MN SMD100LN SMD100TR SMD100NI SMD100CO SMQ621 SMD630 SMQ661 SMQ665A SMQ665B SMQ665C SMQ665D SMQ670 SMQ848 SMQ852Q SMQ852U SMAQUEX2;

run;

\*Question 6 - Importing blood pressure data using SAS Export file;

LIBNAME BPQ XPORT "C:\Users\smecham\Desktop\nhanes\BPQ\_H.xpt";

PROC COPY IN=BPQ OUT=nhanes;

RUN;