**EPID 602**

**Winter 2018**

**Homework 1: Compiling and Cleaning Datasets, Creating Variables**

**Honor Code**

I pledge on my honor that:

I have completed all steps of the attached homework on my own,

I have not used any unauthorized materials while completing this homework, and

I have not given anyone else access to my homework.

Please electronically sign the following honor code below:

Your Name: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Your Student ID: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Signature and date:

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

In the homework assignments we will be using data from the Child Health and Development Studies to answer the question: Is girls’ age of menarche affected by their mothers’ cigarette smoking during pregnancy? The data sets are located in **Canvas**. In the first assignment we will compile and clean the data for our analysis.

Please remember that you may consult with each other while working on this assignment but you must run your own code, write your answers in your own words, and submit your own assignment. Please upload an electronic word document (No PDFs will be accepted) of your assignment with your code pasted in via Canvas before class begins at **1PM on 01/19/18**. Only output that is relevant to the questions should be pasted in below each question. Please do not change the numbering or ordering of the questions.

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

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

3) Your code is well commented (the top of your file should include the homework number and your name, each question should be identified in the code, and each new task should be described by comments) and formatted (indentation and carriage returns should be used to improve readability). 5% will be deducted if either of these two tasks is not completed.

1. Create a permanent library in the Private folder of your IFS space called ***chds***. This is where you will store all of your files (.dat files and SAS files) for later use. Paste a copy of the log output from the LIBNAME statement into your homework. (1 point)
2. The sample for our analysis will be adolescents who participated in the CHDS Adolescent Development Examination. Skim the introduction to the chapter for this exam (Chapter 26) in the CHDS Manual.
   1. According to the manual, how many adolescents participated in the exam? What ages were they? (2 points)

The data file associated with the Adolescent Development Examination is called **adolesnt**. Each observation (row) in this data set corresponds to an adolescent. The column ranges for the variables in the **adolesnt** file start on page 491 of the CHDS Manual.

b. Some of the variable names in the manual contain periods (e.g., C.SEQ.NO and MX.DATE). We will replace the periods with underlines. Why can’t we leave them as periods? (1 point)

1. Now we will start importing data.
   1. Import only the following variables from the **adolesnt** data file (**adolesnt.dat**) using a data step with an INFILE statement. Remember that the underlines in some of the variable names appear as periods in the manual. (2 points)

(Hint: Notice that the record length of this file is long—the last variable is found in column 989. In order to get SAS to read the data in correctly, include the option “LRECL=989” in your INFILE statement.)

WOMAN

PREG

CHILD

AH2\_MENS

AH2\_MENY

AH2\_MENM

AC\_EXAGE

G9\_SEX

Save as a dataset called **teens** in your ***chds*** 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? (1 point)
  2. There are three identifier variables in this file: WOMAN, PREG, and CHILD. Use the CHDS manual to explain what each of these variables represents. (2 points)
  3. Run a PROC PRINT to print only the variables WOMAN, PREG, and CHILD for only the first 30 observations. Notice that there are 2 observations with WOMAN=61. What does this mean? What does it mean when there are 2 observations with the same values of PREG? Why shouldn’t there be 2 observations with the same value of CHILD? (3 points)
  4. Check the **teens** dataset to make sure you do not have any duplicate child identification

1. Information about the girls’ exposure (mother’s smoking during pregnancy) and covariates was collected while their mothers were still pregnant. This information is located in the **basic** data file. Chapter 1 in the CHDS Manual contains information about this file, including column ranges. Each observation corresponds to a child.
   1. Import only the following variables from the **basic** data file (**basicc.dat**) using a data step with an INFILE statement: (2 points)

WOMAN

PREG

CHILD

GRAVAGE

INFANTSX

BIRTHWT

NUMPREG

GRAVEDUC

GRAVWORK

TOTINCOM

AGEMENST

GRAVSMOK

GRAVQUIT

GRAVCIGS

CHLDRACE

Save as a dataset called **basic** in your ***chds*** 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? Why are there so many more observations in the **basic** file than in **teens**? (2 points)

1. One of the covariates for our analysis will be the mother’s age at menarche. The variable AGEMENST from the **basic** file gives this information.
   1. Run PROC FREQ to look at the distribution of the variable AGEMENST and paste the output below. (2 points)

* 1. Without consulting the manual, what is the range of values of AGEMENST? Is this a reasonable distribution of values for the mothers’ ages at menarche? (2 points)
  2. Now use the CHDS manual to check the coding of AGEMENST. Explain how the mothers’ ages at menarche are coded in AGEMENST. (1 point)
  3. How many observations are missing information for the AGEMENST variable? (Hint: This includes both missing values recognized by SAS and observations assigned values for missing or unknown information. Check the manual!) (2 points)

1. Since there are so many observations with missing values of AGEMENST, we will use supplemental information gathered from the mothers’ clinical records. This information is contained in the data file **pnwork**. Each observation in this file corresponds to a pregnancy.
   1. Import only the following variables from the **pnwork** data file (**pnwork.dat**) using a data step with an INFILE statement: (2 points)

WOMAN

PREG

AGEMENAR

Save as a dataset called **pnwork** in your ***chds*** 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? (1 point)
  2. Use the CHDS manual to check the coding of AGEMENAR. Is it the same as AGEMENST? What value is assigned to observations with missing information? Why do you think this value is used instead of 9, which was one of the values for missing information in the AGEMENST variable? (2 points)
  3. Briefly explain a plan for minimizing missing data by combining information from both AGEMENAR and AGEMENST. You do not have to write code for this—just explain what you would do in words (this can be difficult to do clearly!). How would you combine the information? Would you use information from one of the variables preferentially? Which one? Why? What would you do if there is disagreement between the variables? There is no one correct answer to this question. (3 points)

1. We need to merge the **teens**, **basic**, and **pnwork** datasets to create a dataset for our analysis.
2. What identifier variables appear in all 3 datasets? Why is it NOT a good idea for us to simply merge the 3 datasets together using one of these variables? (Hint: Remember that the sample for our analysis is children and refer to question 3c.) (2 points)
3. What procedure do we need to run before we can merge datasets by a common variable? (1 point)
4. Merge the **basic** and **pnwork** datasets together by their shared identifier variable PREG to create a new dataset called **basicpn**. (2 points)
5. How many observations and variables are there in the **basicpn** dataset? (1 point)
6. Now merge the **basicpn** and **teens** datasets by the appropriate identifier variable to create a new dataset called **alldata**. Only retain the observations of children who are in the **teens** dataset**.** (2 points)
7. How many observations and variables are there in the **alldata** dataset? (1 point)
8. Why do we want to retain only observations of children in the **teens** dataset? (1 point)
9. The outcome for our analysis is age at menarche. Run a PROC FREQ of the AH2\_MENS variable to see how many of our adolescents have reached menarche. You do not have to paste the output below.

a. How many have reached menarche? (1 point)

b. Use PROC FREQ to make a 2-way table of the variables AH2\_MENS and G9\_SEX and paste the output below. What does it mean if AH2\_MENS = 9? How many adolescents have this value? Why so many? (3 points)

c. Do you see a problem with one observation in the 2-way frequency table you just made? What is the problem? What do you think we should do about it? (2 points)

1. We want to subset the dataset **alldata** to retain only the girls.

a. There are 2 possible variables in the **alldata** dataset we could use to do this. What are they? Check that the 2 variables agree (i.e., no observations are classified differently by the two variables). Do they? (2 points)

b. Use one of the variables to subset the dataset **alldata** to retain only the girls and name this new dataset **girls**. How many observations are there in the dataset **girls**? (2 points)

c. How many of the girls have nonmissing information about whether they have reached menarche? (1 point)

1. In the **girls** dataset, use PROC UNIVARIATE to get information about the variable GRAVAGE.
   1. What does the variable GRAVAGE represent? (1 point)
   2. Does the distribution look reasonable? Are there any unreasonable values of GRAVAGE? What are they? (2 points)
   3. Recode any unreasonable values and then run PROC UNIVARIATE again. Make a histogram of the distribution of GRAVAGE and paste it below. What are the minimum, maximum, and mean values of GRAVAGE? (2 points)

In questions 11-13 we will create our outcome and exposure variables in the **girls** dataset.

1. The outcome for our analysis is age at menarche of the adolescent girls. We will create two measures of age at menarche: a continuous variable and a binary variable for late menarche (defined as age 14 or older). In order to construct these new variables, we will need to use the variables AH2\_MENS, AH2\_MENY, and AH2\_MENM. Familiarize yourself with the distributions of these variables (e.g., using PROC FREQ).
2. Remember that in the manual these variables are named AH2.MENS, AH2.MENY, and AH2.MENM. Using the CHDS manual for reference: (2 points)

* Recode missing values of these variables appropriately. For our purposes, we will consider Unknown, Skipped, and Not Interviewed values all as missing.
* Use a LABEL statement to assign labels to these variables that are meaningful to you.

1. There are two issues we must address before creating our continuous variable. First, there are instances where girls have reached menarche and have information for AH2\_MENY but are missing information for AH2\_MENM. How many instances? Recode them such that their AH2\_MENM=0. Second, there instances where girls have AH2\_MENS=0 but have information for AH2\_MENY and AH2\_MENM. How many instances? We will recode these girls such that AH2\_MENS=1 (1 point).
2. Create the continuous measure of age at menarche. Name the variable TEENMENS and incorporate the following: (2 points)

* Combine the AH2\_MENY and AH2\_MENM variables to create a continuous measure of age at menarche in years. (Hint: This variable will contain decimal values because of the months information.)
* Code girls who have not reached menarche yet as missing for this variable. Notice that this means analyses using this variable only apply to girls who have reached menarche by the time of the interview.
* Use a LABEL statement to assign a label to the variable TEENMENS that is meaningful to you.

1. Run a PROC UNIVARIATE to get information about the variable TEENMENS, including a histogram. Paste the histogram below. How many girls have a nonmissing value for this variable? What is the mean age at menarche among girls who have reached menarche? (2 points)
2. How might our analysis be affected by the fact that we coded girls with missing AH2\_MENM values as if they had AH2\_MENM=0? What is an advantage of doing this? What is a disadvantage? Do you think this was a good idea? If so, why? If not, how would you have handled these observations? (3 points)
3. Similarly, how might our analysis have been affected by recoding AH2\_MENS from 0 to 1 for girls with AH2\_MENY and AH2\_MENM values? What might have caused this inconsistency? Do you agree with how we handled these observations? Why? (3 points)
4. Please create another new variable in the **girls** dataset.
5. Create the binary measure of late age at menarche. Name the variable LATEMENS. Incorporate the following: (2 points)

* Assign LATEMENS=1 to girls who HAD reached menarche at the time of the adolescent interview, and reported reaching menarche when they were aged 14 years old or older.
* Assign LATEMENS=0 to girls who HAD reached menarche at the time of the adolescent interview, and reported reaching menarche when they were younger than 14 years old.
* Assign LATEMENS=1 to girls who HAD NOT reached menarche yet at the time of the adolescent interview if they were 14 years or older at the time of the interview. (Hint: The variable AC\_EXAGE gives the age at the time of the adolescent interview.)
* Assign missing values to girls who HAD NOT reached menarche yet at the time of the adolescent interview if they were under the age of 14 at the time of the interview.
* Use a LABEL statement to assign a label to the variable LATEMENS that is meaningful to you.

1. Why do you think we are assigning LATEMENS=1 to girls who had not reached menarche yet at the time of the adolescent interview if they were aged 14 years or older at the time of the interview? (1 point)
2. Run a PROC FREQ of the variable LATEMENS and paste the table below. (1 point)
3. The exposure for our analysis is maternal smoking during pregnancy, defined as a categorical measure of the number of cigarettes the mother smoked per day. In order to construct this new variable, we will need to use the variables GRAVSMOK and GRAVCIGS. Familiarize yourself with the distributions of these variables (e.g., using PROC FREQ).
4. Using the CHDS manual for reference: (2 points)

* Recode missing values of these variables appropriately. For our purposes, we will consider Unknown (including unknown quantity), Skipped, and Not Interviewed values all as missing.
* Use a LABEL statement to assign labels to these variables that are meaningful to you.

1. Create the categorical measure of maternal smoking during pregnancy. Name the variable MOMCIGS and incorporate the following: (2 points)

* Mothers who reported that they never smoked; smoked until the current pregnancy; or once smoked, but now do not, should be assigned to the zero category regardless of how many cigarettes they smoked per day when they used to smoke or whether they have missing information about how many cigarettes they used to smoke per day.
* Mothers with missing information about whether they smoke should be assigned missing values.
* For mothers who reported during pregnancy that they currently smoked, the variable should have 4 categories corresponding to zero cigarettes per day, 1–9 cigarettes per day, 10–19 cigarettes per day, and ≥ 20 cigarettes per day.
* Use a LABEL statement to assign a label to the variable MOMCIGS that is meaningful to you.

1. Run a PROC FREQ of the variable MOMCIGS and paste the table below. (1 point)
2. Given the constraints of the CHDS dataset, do you think this was a good way to define our exposure variable? Why or why not? What were some other possibilities? (2 points)
3. Save the dataset **girls** as a permanent dataset in the library ***chds***. Verify that the dataset appears in the correct folder on your IFS space. Paste a copy of the log output from your DATA statement below. (1 point)