

# Fall 2011 BioEpi 691F: Practical Data Management and Statistical Computing

## Assignment 8: Using SET and MERGE statements

Due: Tuesday  
22 NOV 2011

In this assignment you will be using SET and MERGE statements along with the automatic indicator variables: (IN variables, and FIRST. and LAST. variables) to join datasets and/or make appropriate subsets.

You should have copies of most of the data and program files that you need for this assignment saved from the previous assignment. If you had problems with assignment #7, you will need to correct errors (preferable) or run a copy of the programs that are attached to the solutions page. All the programs and data files are still available on the website, should you need new copies of the data.

### Reading

#### Course Notes:

- SAS Notes Unit 3

- 
- I. Starting with the formatted, labeled version of the burner data from assignment #7, do the following:
    - **NEW VARIABLES and RECODING**
      - a. Create a new variable for height in inches, dropping the separate variables for the feet and inches parts of height from the data set.
      - b. Recode any unexpected or unusual values you found for CLASS and HOWMANY as missing values.
      - c. For all players who indicated "no burners" last year, recode the value of HOWMANY from missing to zero.
      - d. Create an indicator variable to identify Freshman that takes the value 1 for freshman and zero for upperclassmen (sophomore/junior/senior/5th yr).
  - II. Label the new variables you created. Assign formats as appropriate.  
(Tip: add to your format program from the previous assignment, as needed to create new formats, and rerun it).
  - III. Print the first 20 observations in the file using the labels, and run contents on your dataset.  
Note: the **obs=20** option after the data set name will limit the number of observations printed. *DO NOT TURN IN A LIST OF ALL 700+ OBSERVATIONS!*  
(Use: **PROC PRINT DATA=libn.dsn(obs=20) labels;** )

- **SUBSETTING**

1. **Create 2 temporary datasets** from the burner data, one with those who reported that they **never had burners**, and one with those who have **ever** had burners. Keep only subject id, freshman indicator, and number of burners last season in these data sets.
2. For each data set, produce frequency tables of freshman vs. upperclass status, and number of burners last year. Use appropriate titles.

- **CONCATENATING**

1. Rejoin (concatenate) the never/ever burner data sets into a single data set. In the process of rejoining, use **IN** variables to create a new variable to indicate ever/never had a burner in this new data set.
2. Print frequency tables of freshman/upperclass status and ever/never had burners using your re-joined data.

---

II. The file [binj1.sas7bdat](#) contains a small subset of variables from the injury report forms. These are reports of burners that occurred during the prospective study period. Write a program to:

- a. Look at **contents** of this file.  
How many burners were reported in total during the prospective study?
- b. **MERGE** the injury file with the initial survey file, **matching by player id**.

Look at the contents of the new merged file. How many observations are in the data set? How many observations were in the initial survey data set? What does this tell you?

- c. Create 2 subsets from the merged file,  
(1) those with at least one injury (found in binj1) reported **during the study**, and  
(2) those who did not report any injuries during the study year.

(Note: last and howmany refer to burner injuries **a history of burners before** the study year -- **only the binj1 data file has the burners reported during the study year**).

How many players did not have any burners in the study year?  
How many players had at least one burner during the study?

Look at your program for part **b**.

How can you create a variable, **as you merge** the injury and initial survey files, to indicate "current year burners" vs. none? Think about using IN variables and/or FIRST. and LAST. Variables.

---

III. **Optional:**

This problem is designed to demonstrate the purpose of the RETAIN statement.

- a. Write a program that will input the data for the data set MULT (under "LIST OF SCORES BY STATUS: NEW VERSION") on page 3.17 of the class SAS notes. Then use the program demonstrating the use of the RETAIN statement to re-organize the data with one observation per subject.
- b. Try it without the RETAIN statement to see the difference. Print out the new dataset each time. Use appropriate titles.
- c. Write a program to recreate MULT (under a new name) from the one observation per subject data set. Run this program.

---

IV. **Extra Credit:**

The merged data file you created in step II has multiple observations on some players -- the players who reported more than one injury during the season.

From the merged file, create a new file that has one observation per player, and contains a count of the number of burners reported during the study year: BCOUNT. This should have value zero if the player did not report any burners.

Hint: use a RETAIN statement to hold onto the BCOUNT variable; use **first.** and **last.** to identify the start and end for a player.

---

Please remember to copy all the files you want to save to your own disk.

**Please hand in copies of your log and edited output for each problem.**

Use comments to make the program steps clear, and use titles that give the problem number and letter, as well as descriptive titles on the output.

By the phrase '**edited output**', I expect that you save your output files (or copy and paste results from the output window into WORD), and edit the output in WORD (or another word processor) so that spacing, font style and size, page-breaks, etc. are reasonable.

For example:

- Avoid tables where the end of the lines wrap to the next line -- change your fonts and/or margins and/or page orientation (portrait/landscape) to fix this if it is a problem
- Do not split a table or list across pages, unless the table is too long to fit on 1 page.
- Number pages and use footnotes or headers to include your name.
- Number tables throughout the document. This can be within problem number or sequentially through the whole document.
- Respond to any questions in the assignment on the output page.

---

If you are working in a computer lab, please erase your files from the hard drive and log off when you leave the computer lab.

---