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

Assignment 7: Documenting a Data Set **Due: Tuesday 15 NOV 2011**

In this assignment you will read a SAS data set containing a subset of the variables from the **Initial Survey data of the Burner Study**. You will save a new version of the data after adding labels and assigning formats to the variables, and identifying and recoding missing and invalid data. Finally you will produce a summary tables on a few of the variables.

Reading:

Course Notes: Introduction to SAS for Data Management Units 1, 2, and 3

Course Notes Part I: Section 3.4 (Burner Study)

Additional Reference: SAS Language Reference: Concepts, Version 9.2.

Save the data file **burner1.sas7bdat** to an appropriate working directory.
Print the coding manual for the data – the last page of this document.

-
1. Before you make any changes to the data, take a look at what is in the file:
 - a. Use PROC CONTENTS to list information on the burner1 SAS data set. This file contains data from the initial survey of the burner study.
 - b. Create frequency tables (use PROC FREQ, Unit 5) for the variables CLASS, EVER, LAST, HOWMANY, NOTREPORT, EQUIP

Copy and paste your log and output files to a WORD document; edit for pagination, readability and print to hand in.

-
2. Write another program to create a format file for the data.
 - a. Use PROC FORMAT to create a format file for your data using information from the coding manual to define the formats. Begin the format procedure with the statements:

```
LIBNAME libref 'drive:\folder';  
PROC FORMAT CNTLOUT=libref.filename;
```

where:

libref is the name on your LIBNAME statement
drive:\folder specifies the location for storing your format file
filename is the name of the format file you will create and save – **don't use the same name as your data file!!!**

Note that PROC FORMAT with the CNTLOUT option creates a special type of SAS data file that stores format information.

At the end of the format procedure, after the RUN; statement, include the following statements. These statements print out your format list in a user-friendly manner – e.g., they will help document your formats.

```
** copy format file into a temporary file called FCODES **;  
** and keep format name, startvalue and format label **;  
data fcodes(keep=fmtname start label);  
    set libref.filename;  
run;  
  
** sort new file by format names **;  
proc sort data=fcodes;  
    by fmtname;  
run;  
  
** print list of format names nicely **;  
** using label option of print **;  
proc print data=fcodes label;  
    by fmtname;  
    id fmtname;  
    var start label;  
run;
```

Save your format program, run it, and save the output and log. Copy and paste the log and output to WORD (edited into a single document) and print to hand in. Pay attention to page breaks.

3. Write a program to create a new version of the burner data set that will:

- a. Include a program header to describe the purpose of the program, along with documentation information such as input and output data files.
- b. Add an OPTIONS statement to control pagesize, linesize, suppress printing of page numbers and date, and turn off centering.
- c. Read in the formats you created in step 2, using the following statements:

```
PROC FORMAT CNTLOUT=libref.filename;  
run;
```

where *libref* is the libname for the directory where your format file is stored.

Note: CNTLOUT **writes** (creates) a format file, while CNTLIN **reads** a format file you have already saved.

- d. In the DATA step to create your new version of the data:
 - i. Add a LENGTH statement to control variable length. For numeric codes a length of 3 is adequate. You decide what length to use for other numeric variables. Note that the "default=" option will not change lengths of variables that already have an assigned length.

- ii. Add a LABEL statement to label variables with descriptive labels.
- iii. Use a FORMAT statement in the data step to assign the formats you created in step 2 to the appropriate variables.
- iv. Recode missing values as applicable, to SAS missing values.
- e. Use PROC FREQ create frequency tables for the same variables as in step 1, this time using your labeled, formatted file.
- f. Use PROC CONTENTS to look at the new dataset information.

Include TITLES and/or FOOTNOTES in your program. Use comments throughout your program.

Copy and paste the log and output to a WORD document and edit for pagination, readability. Hand in copies of the log and edited output files.

Comment on differences you note between the results of the PROC CONTENTS and frequency tables from parts 1 and 3.

Burner Study Coding Manual for HW7

The baseline survey data are contained in the file **burner1.sas7bdat**. The data correspond to the following variables:

Variable	Description	Codes/Value Range	Missing Code
pid	Player ID	0001-9999 of form Nnnn, digit N indicates college code, nnn player within college	9999
htft	feet part of height in ft, in	5-6	9
htin	inches part of height in ft,	0-11	99
weight	weight in pounds	3 digit numeric	999
class	college class	1=Freshman 2=Sophomore 3=Junior 4=Senior 5=5th year	9
ever	ever had burner in past	1=yes, 0=no	9
last	had a burner last season	1=yes, 0=no	9
howmany	# of burners last season	numeric count: 0-98	99
notreport	ever failed to report burner	1=yes, 0=no	9
equip	wear protective equipment	1=yes, 0=no	9