

Using SDA on the Web to Extract Data from the General Social Survey and Other Sources*

Brett Presnell
Dept. of Statistics
University of Florida

March 19, 2001

1 Getting the Data

The *SDA: Survey Documentation & Analysis* web site [2] is a set of programs “for the documentation and Web-based analysis of survey data.” This site includes data from several different surveys, including the General Social Survey (GSS). For further information on the GSS specifically, you can also visit the GSS web site [1].

From the SDA home page, click on the “SDA Archive” button. This will take you to a listing of the surveys available. You may wish to examine some of these as possible sources for data, but we will focus on the GSS, so now click on the “GSS Cumulative Datafile 1972-1996” link. This will take you to the *SDA Demonstration Survey Data Archive* page (see Figure 1

At this stage it is probably worth while to click on the “Extra Codebook Window” button, which will open another browser window for the *codebook* for the GSS. This codebook is your guide to the variables included in the GSS and their codings. Clicking on the “Group Headings” heading will allow you to examine the variables grouped according to the type of topic addressed. This is useful when looking for variables that might be of interest. Once you know the names of the variables that you are interested in, it may be more efficient to find them again through the alphabetically ordered list. The appendices may also be useful at times. In particular Appendix U (Variables by Year) indicates in which years each variable was measured (although this does not seem to include the 1996 GSS).

After you find the variables that you want to analyze, return to the *SDA Demonstration Survey Data Archive* page, choose “Download a customized subset of variables/cases”, and click on the “Start” button. This takes you to the *SDA Customized Subset of Variables/Cases* page (see Figure 2. Possibly the first thing you will want to do here is to choose “Blank” as the delimiter for your data file, as opposed to “None”. You will want to create both a data file and a codebook, so leave those selections alone.

For our example, we have decided to get the variables sex (respondents sex), age (age of respondent), income (total family income), educ (highest year of school completed), owngun (have gun in home), hunt (does respondent or spouse hunt), and gunlaw (favor or oppose gun permits) for the 1996 GSS. Thus we choose year(96) as our filter variable, and enter the other variables as individual variables to include. Note that you can use all sorts of filters here to narrow the data, and you may even wish to use this facility to eliminate observations that have missing or bad data on the

*© Copyright 2000 by Brett Presnell

variables that you wish to use. Here we will take the approach of doing the latter kind of filtering in SAS instead.

Once you have entered all your variables click on the “Continue” button at the top of the page. This will take you to a page for checking your subset specifications. If everything looks right, click on the “Create the Files” button, which takes you to the *Download Files* page 3.

At this stage you may want to first click on the “Data file” and “Codebook” links to make sure that everything looks reasonable before downloading these files. If you do you will see that the first two variables are the year of the survey and the case id of the observation. These are always downloaded by default, and are not separated by a blank. Here are the first and last few lines of our data file:

```
96  1 1 79 13 12 2 4 1
96  2 1 32 13 17
96  3 1 55 13 18 2 4 1
... lines deleted ...
962902 1 40  9 12 2 4 2
962903 1 36 12  9 2 4 2
962904 2 33 12 12
```

The codebook will tell you what variables are in your data file, what columns they are entered in, and how they are coded. This is essential information for analyzing your data. Thus for example, in our data file, the codebook entry for income gives

```
income                TOTAL FAMILY INCOME
```

```
    In which of these groups did your total family income,
    from all sources, fall last year before taxes, that
    is? Just tell me the letter.
```

```
VALUE  LABEL
    0  NAP
    1  LT $1000
    2  $1000 TO 2999
    3  $3000 TO 3999
    4  $4000 TO 4999
    5  $5000 TO 5999
    6  $6000 TO 6999
    7  $7000 TO 7999
    8  $8000 TO 9999
    9  $10000 - 14999
   10  $15000 - 19999
   11  $20000 - 24999
   12  $25000 OR MORE
   13  REFUSED
   98  DK
   99  NA
```

```
Data type: numeric
Missing-data codes: 0,98,99
```

Record/columns: 1/13-14

Note in particular that this variable is entered in columns 13–14 of the data file, and that only codes 1–12 actually correspond to a meaningful income level. To keep things simple, we will filter out observations with other codes, at least if we decide to use income in our analysis. The codebook for our variables is included in this document in Appendix A for reference.

If you have not already done so, you should download the data and codebook files now and save them to disk.

2 Using SAS to Read and Analyze the Data

Attached at the end is a SAS program for reading these data. The main things to note are:

1. The delimiters `/* . . . */` are used for comments in the SAS program.
2. The column numbers from the codebook are used in the `input` statement to read the data.
3. Various if/then/else statements are used to delete observations with bad or missing data. Note that this is not necessarily the best thing to do (missing data can be important), but it will do for purposes of this project.
4. `income` is recoded into scores.
5. `gunlaw` is coded so that 1 represents “yes” and 0 “no.” This is necessary since we plan to use this variable as a response in a logistic regression. We have also done the same with `owngun`, but this is not necessary unless we also plan to use this variable as a response.
6. `PROC PRINT` can be useful to see the SAS data set that you have created in order to be sure that everything looks right. After you have everything settled though you will probably want to comment out or delete these lines since the printout will be very long otherwise.
7. Summary statistics can be produced by various SAS procs. You do not necessarily have to use these, but in case your interested, I’ve added links to their manual pages. You might also just want to look at the *SAS Procedures* manual.

References

- [1] National Opinion Research Center. General Social Survey. <http://www.icpsr.umich.edu/GSS99/>, March 1999.
- [2] Berkeley Computer-assisted Survey Methods Program (CSM), University of California. *SDA: Survey Documentation & Analysis*. <http://csa.berkeley.edu:7502>, March 2000.

A Codebook

CODEBOOK

1972-1996 General Social Survey Cumulative File

CONTENTS

item		page
CASEID	Case identification number	1
sex	RESPONDENTS SEX	1
age	AGE OF RESPONDENT	1
income	TOTAL FAMILY INCOME	2
educ	HIGHEST YEAR OF SCHOOL COMPLETED	2
owngun	HAVE GUN IN HOME	3
hunt	DOES R OR SPOUSE HUNT	3
gunlaw	FAVOR OR OPPOSE GUN PERMITS	4

1972-1996 General Social Survey Cumulative File Page 1

CASEID Case identification number

Data type: character
Record/columns: 1/1-6

sex RESPONDENTS SEX

VALUE	LABEL
1	MALE
2	FEMALE

Data type: numeric
Record/column: 1/8

age AGE OF RESPONDENT

What is your date of birth?

VALUE	LABEL
98	DK
99	NA

Data type: numeric
Missing-data codes: 0,98,99
Record/columns: 1/10-11

income TOTAL FAMILY INCOME

In which of these groups did your total family income, from all sources, fall last year before taxes, that is? Just tell me the letter.

VALUE	LABEL
0	NAP
1	LT \$1000
2	\$1000 TO 2999
3	\$3000 TO 3999
4	\$4000 TO 4999
5	\$5000 TO 5999
6	\$6000 TO 6999
7	\$7000 TO 7999
8	\$8000 TO 9999
9	\$10000 - 14999
10	\$15000 - 19999
11	\$20000 - 24999
12	\$25000 OR MORE
13	REFUSED
98	DK
99	NA

Data type: numeric
 Missing-data codes: 0,98,99
 Record/columns: 1/13-14

educ HIGHEST YEAR OF SCHOOL COMPLETED

What is the highest grade in elementary school or high school that you finished and got credit for?

VALUE	LABEL
97	NAP
98	DK
99	NA

Data type: numeric
 Missing-data codes: 97,98,99
 Record/columns: 1/16-17

owngun HAVE GUN IN HOME

Do you happen to have in your home (IF HOUSE: or

B SAS Program

```
options nodate nocenter linesize=64;

data gunctrl;
  infile 'guncontrol.txt';
  input sex 8 age 10-11 income 13-14 educ 16-17
        owngun 19 hunt 21 gunlaw 23;
  /* deleting observations with bad or missing data */
  if age>=98 then delete;
  if income=0 or income>=13 then delete;
  if educ>=97 then delete;
  if owngun=0 or owngun>=3 or owngun=" " then delete;
  if hunt=0 or hunt>=8 or hunt=" " then delete;
  if gunlaw=0 or gunlaw>=8 or gunlaw=" " then delete;
  /* convert income to scores representing income in $1000's */
  if income=1 then income=0.5;
  else if income=3 then income=3.5;
  else if income=4 then income=4.5;
  else if income=5 then income=5.5;
  else if income=6 then income=6.5;
  else if income=7 then income=7.5;
  else if income=8 then income=9;
  else if income=9 then income=12.5;
  else if income=10 then income=17.5;
  else if income=11 then income=22.5;
  else if income=12 then income=30;
  owngun = 2-owngun; /* 1=yes, 0=no */
  gunlaw = 2-gunlaw; /* 1=yes, 0=no */
  n=1; /* for logistic regression of ungrouped data */

/*
proc print;
run;
*/

proc sort data=gunctrl out=sexsort;
  by sex;
run;

proc means data=sexsort;
  by sex;
  var income;
run;

proc univariate data=sexsort;
  by sex;
  var income;
  output q1=q1 median=med q3=q3;
run;

proc freq data=gunctrl;
  tables sex*owngun*gunlaw / nopercnt nocol norow;
run;

proc genmod data=gunctrl; class owngun hunt;
  model gunlaw/n = income educ owngun hunt
    / dist=bin link=logit;
run;
```

Figure 1: SDA Demonstration Survey Data Archive web page.

Figure 2: Selecting variables.

Figure 3: Downloading the data set and codebook.