Using SPSS to Reformat Data Records from One to Several

Last revised: 08/27/99

Researchers frequently wish to analyze secondary data in a format that is different from the publicly released version. In the following example, data from a roster of household members is reformatted to separate records for each child aged 0 to 17 years. The data are taken from the National Survey of Families and Households, Wave 2. The household roster included information on each person in the household:

- Name, sex, age, marital status and whether married before, relation to R
- Whether each of R's biological children is also a biological child of spouse/partner
- Education, working full or part time, enrolled in last four months and school type, economic situation, number of children (and number living here)
- If absent spouse: reason for absence.

The variables used in this example are:

MA1	Main respondent's ID number (5 digits)
MA7	Main respondent's sex (1=Male, 2=Female)
MA8	Main respondent's age (in years)
MB1NUM	Number of additional people in household
MB4Pxx	Sex of household member (1=Male, 2=Female, xx ranges 1 to 16)
MB5Pxx	Age of household member (in years, xx ranges 1 to 16)

The first three cases look like:

MA1	MA7	MA8	MB1NUM	MB4P01	MB5P01	MB4P02	MB5P02	MB4P03	MB5P03	MB4P04	MB5P04	MB4P05	MB5P05	MB4P06-16	MB5P06-16
3	1	56	5	2	58	1	39	1	22	1	9	1	10	9	99
11	2	85	0	9	99	9	99	9	99	9	99	9	99	9	99
29	2	62	2	2	6	2	5	9	99	9	99	9	99	9	99

where all data for the 6th through 16th household members are missing (MB4Pxx (sex) is coded 9 and MB5Pxx (age) is coded 99).

The result of reformatting the data will look like:

MA1	MA7	MA8	MB1NUM	SEX	AGE
3	1	56	5	1	9
3	1	56	5	1	10
29	2	62	2	2	6
29	2	62	2	2	5

where there is one record for each household member aged 0 to 17 years that includes selected characteristics of the main respondent. In the example shown, there are two records each for main respondent (MA1) 3 and 29. Main respondent (MA1) 11 lives alone (MB1NUM = 0). Note that for main respondent (MA1) 3 only 2 of the 5 additional household members are aged 0 to 17 and included in the reformatted data. Both additional household members are in this age group for main respondent (MA1) 29 and included in the reformatted data.

SPSS can handle this data in two ways. The first method uses vectors and loops to repeatedly process the data for each household member. The second method uses the repeating data facility to read the data set in the desired format.

Vectors and Loops

The following SPSS program uses vector and loops in an input program to process the roster data and reformat the data. The vectors serve as shortcuts to the lists of personal characteristics that are repeated for each household member and the loops control the processing of data for each household member in turn.

```
file handle  nsfh2 / name= '/usr/ftp/pub/nsfh/drmain.007'
lrecl=8231.
input program.  Ø
data list file=nsfh2 /
 MA1       0001-0005
 MA7       0016
 MA8       0017-0019
 MB1NUM    0021-0022
 #MB4P01    0026 Ù
 #MB4P02    0050
 #MB4P03    0074
 #MB4P04    0098
 #MB4P05    0122
 #MB4P06    0146
 #MB4P07    0170
 #MB4P08    0194
 #MB4P09    0218
 #MB4P10    0242
 #MB4P11    0266
 #MB4P12    0290
 #MB4P13    0314
 #MB4P14    0338
```

```
 #MB4P15    0362
 #MB4P16    0386
 #MB5P01    0027-0028
 #MB5P02    0051-0052
 #MB5P03    0075-0076
 #MB5P04    0099-0100
 #MB5P05    0123-0124
 #MB5P06    0147-0148
 #MB5P07    0171-0172
 #MB5P08    0195-0196
 #MB5P09    0219-0220
 #MB5P10    0243-0244
 #MB5P11    0267-0268
 #MB5P12    0291-0292
 #MB5P13    0315-0316
 #MB5P14    0339-0340
 #MB5P15    0363-0364
 #MB5P16    0387-0388
.
leave ma1 ma7 ma8 mb1num.  Ú
vector vsex  =  #MB4p01 to #MB4p16.  Û
vector vage  =  #MB5p01 to #MB5p16.

loop #i=1 to 16.  Ü
compute          sex = vsex(#i).  Ý
compute          age = vage(#i).

end case.  Þ
end loop.  Ü
end input program.  Ø
execute.

* select records for hh member between age 0 and 17 (other than
R).
select if (age le 17).  ß
execute.
```

Ø SPSS uses the Input Program – End Input Program commands to allow users to read many types of file complex structures. In this case it will allow us to read in the single long record for a household and write out separate records for each household member with the household and main respondent information attached. Every Input Program command must match an End Input Program later in the program.

Ù While this Data List is quite basic, I have chosen to read the characteristics of household members in as scratch variables. This is defined by appending the # to the front of the variable name. Scratch variables are not written to the system file and are not available to procedures. This means that they can be used to construct regular variables in the data set, but will not take up unnecessary disk space.

Ú The Leave command designates that the variables listed should be written to each SELECT record when read from a single input record. In this case, the variables

represent the household or main respondent characteristics that will be attached to each record for the other household members.

Û The Vector command defines a nickname for a list of variables. The vector names (vsex and vage in this example) cannot be the same as any variable name used in the data set. This example uses the "v" prefix on the sex and age vectors, so that variables named "sex" and "age" can be created below. The list of variables that make up the vector follow the equals (=) sign. Note that the "to" designation on the variable list requires that the variables making up the vector be consecutive on the active file. The Data List used was written with this in mind so that all of the household member sex variables (#MB4P01 through #MB4P16) were read and then the household member age variables (#MB5P01 through #MB5P16) were read EVEN THOUGH THE COLUMNS IN THE RAW DATA WERE NOT IN THIS ORDER.

Ü The Loop – End Loop commands define the looping process where all of the characteristics (here, sex and age) are processed for each household member listed in turn. An index variable, #i, is used to refer to the vector elements and counts the number of times the loop is executed. Again, the # designates that the index variable is a scratch variable that will not be written to the SELECT data file. The loop is executed once for each member of the household (other than the main respondent) based on the value of MB1NUM. Note that for main respondent 11 MB1NUM = 0 and this Loop statement evaluates to:

```
loop #i=1 to 0.
```

In this case, the loop is not executed and SPSS does not generate any errors or warnings.

If the data did not contain an indicator variable for the number of household members (MB1NUM), this command could have been written:

```
loop #i=1 to 16.
```

and the loop would have been executed 16 times for every case. While this would work in this situation, it is less efficient than limiting the number of times the loop is executed to the number of eligible household members.

Every Loop command must match an End Loop command later in the program.

Ý Here two new variables are computed, sex and age, for the current household member. The variables are equal to the current value of the vector for the current iteration of the loop. This will be shown in more detail below.

Þ The End Case command signals that the case is complete and immediately passes control for the case out of the input program.

ß The Select If commands indicates that only records for household members aged 0 (less than one year) to 17 years should be used in analyses or saved, depending on commands that follow.

Let's take the first case, main respondent 3, and track the looping process. For this record MB1NUM = 5, thus, the Loop command is evaluated as:

loop #i=1 to 5.

so the loop is executed five times.

When #i=1:

When i=1:
ASEX{1} = MB4P01 SEX = 2
AAGE{1} = MB5P01 → AGE = 58 → OUTPUT = No.

When i=2:
ASEX{2} = MB4P02 SEX = 1
AAGE{2} = MB5P02 AGE = 39 → OUTPUT = No.

When i=3:
ASEX{3} = MB4P03 SEX = 1
AAGE{3} = MB5P03 AGE = 22 → OUTPUT = No.

When i=4:
ASEX{4} = MB4P04 → SEX = 1
AAGE{4} = MB5P04 → AGE = 9 → OUTPUT = Yes.

When i=5:
ASEX{5} = MB4P05 SEX = 1
AAGE{5} = MB5P05 → AGE = 10 → OUTPUT = Yes.

After the loop is executed the designated number of times the next record is read from the input file and the program continues until all the input records are processed. This results in a data file like the one shown at the bottom of page 1.

Repeating Data

The following SPSS program uses another Input Program with a Repeating Data command to read the characteristics for household members and create separate records.

```
file handle  nsfh2 / name= '/usr/ftp/pub/nsfh/drmain.007'
lrecl=8231.
input program.  Ø
data list file=nsfh2 / Ù
 MA1       0001-0005
 MA7       0016
 MA8       0017-0019
 MB1NUM    0021-0022
```

```
.
repeating data starts=24 / occurs=mb1num / Ú
     data = person    1-2
            sex       3
            age       4-5
            junk      6-24 (A).

end input program.  Ø
execute.

* select records for hh member between age 0 and 17 (other than
R).
select if (age le 17).  Û
execute.
```

Ø SPSS uses the Input Program – End Input Program commands to allow users to read many types of file complex structures. In this case it will allow us to read in the single long record for a household and write out separate records for each household member with the household and main respondent information attached. Every Input Program command must match an End Input Program later in the program.

Ù The Data List command reads in the variables common to all household members. Here, it is the main respondent's ID, sex, and age and the number of other household members.

Ú The Repeating Data command reads input cases whose records contain repeating groups of data. For each repeating group, Repeating Data builds one output case in the active system file. In effect, Repeating Data generates a Leave command that references all previously defined variables (see **Ù**).

The Starts subcommand indicates the beginning location of the repeating data segment. Starts is required and can specify either a number or a variable name.

The Occurs subcommand specifies the number of repeating groups on each input case. Occurs is required and can specify a number if the number of groups is the same on all input cases or a variable if the number of groups varies across input cases. In this example, MB1NUM indicates the number of other household members and varies across input cases, so we indicated `occurs=mb1num`. However, since every record does in fact include fields for 16 other household members, we could have specified `occurs=16`.

The Data subcommand specifies a variable name, location within each segment, and format type (if necessary) for each variable to be read from the repeating groups. The Data subcommand is required and must be the last subcommand on the Repeating Data command. Note that the column locations are relative to the Starts= location and the total number of columns of the repeating data segment. In the example, we are only interested in the sex and age of other household members, so all the rest of their characteristics

were read into a single variable, called JUNK. We would want to drop JUNK if we saved the file to disk to save diskspace. However, it is necessary to read all the columns of the repeating data segment so that the Input Program actually reads the correct columns for all household members in the repeating data segment.

Û The Select If commands indicates that only records for household members aged 0 (less than one year) to 17 years should be used in analyses or saved, depending on commands that follow.

Recommended Sources

SPSS Reference Guide

SSCC has a complete set of SPSS documentation. These documents are circulated by the CDE Print/Virtual Library in 4457 Social Science.