Session 4 - SPSS[1]

Using a real social science data set. Matching with renaming. Calculating new variables using Do Repeats. Selecting a subset of cases.

Today we will start applying what we have learned about SPSS on a real social science data set. Charles Hirschman has given me permission to use his data for this purpose. The documentation can be found at: http://csde.washington.edu/csde/vietnam/documents.html

and the data can be found at: http://csde.washington.edu/csde/vietnam/data.html

We will use the 1995 data.

The data are downloadable as zipped SPSS portable files. SPSS can import SPSS portable files with the following program:

```
*** PROGRAM BEGINS HERE *** RUN IF NOT WORKING IN Savery 112.
* hh95vls1.SPSS household data.
title 'c:\all\SPSSclass\vn95\progs\hh95vls1.SPSS' .
* Documentation and data from following sites .
* http://csde.washington.edu/csde/vietnam/documents.html .
* http://csde.washington.edu/csde/vietnam/data.html .
* This program converts an spss portable file to SPSS.
IMPORT
  FILE='C:\all\sasclass\VN95\data\hh95vls1.por'.
SAVE OUTFILE='C:\all\spssclass\vn95\data\hh95vls1.sav'
  /COMPRESSED.
execute .
*** PROGRAM ENDS HERE *** .

*** PROGRAM BEGINS HERE *** RUN IF NOT WORKING IN Savery 112.
* i95vls1.SPSS individual level data .
title 'c:\all\SPSSclass\vn95\progs\i95vls1.SPSS' .
* Documentation and data from following sites .
* http://csde.washington.edu/csde/vietnam/documents.html .
* http://csde.washington.edu/csde/vietnam/data.html .
* This program converts an spss portable file to SPSS.
IMPORT
  FILE='C:\all\sasclass\VN95\data\i95vls1.por'.
SAVE OUTFILE='C:\all\spssclass\vn95\data\i95vls1.sav'
  /COMPRESSED.
execute .
*** PROGRAM ENDS HERE *** .
```

As with some of our practice programs, there are separate files for household and person records. This file is a bit different from the files we have worked with. There is one household record for each household. One person in the household was the informant for the household questions, and also was interviewed for the person level data. Multiple people in the household were also questioned about him or herself. Each person was also asked some questions about other people in the household.

Another difference is that there is not a single variable that uniquely identifies each household. Instead, two variables must be used, and these have different names in the two files. Both of these variables must be used to uniquely identify a household. And, the variables must be renamed so that the variable names are common to both files.

	Household	Individual

[1]This document has been prepared by Patty Glynn, University of Washington.

```
                                   Survey          Survey
Commune ID                         h5q2            i5q2
Household ID (unique in Commune)   h5q7            i5q7
```

The following program will rename the variables, and match the two data sets.

```
*** PROGRAM BEGINS HERE *** .
* vn95.SPSS .
title 'c:\all\SPSSclass\vn95\progs\vn95.SPSS'.
* Documentation and data from following sites .
* http://csde.washington.edu/csde/vietnam/documents.html .
* http://csde.washington.edu/csde/vietnam/data.html .
** Put Household and Individual Level Data together .
** Store SPSS data set with all variables and cases .

match files
  / table = 'c:\all\SPSSclass\vn95\data\hh95vls1.sav'
     / rename h5q2 = commid h5q7 = hhid  /
  / file = 'c:\all\SPSSclass\vn95\data\i95vls1.sav'
     / rename i5q2 = commid i5q7 = hhid  /
  / by commid hhid
  / map .
display labels .
save outfile = 'c:\all\SPSSclass\vn95\data\vn95.sav' .
execute .
```

Matching is a tricky business. If a match does not go well, you will append variables to the wrong cases, and all of your results will be wrong. Of course it is ALWAYS important to check your log file carefully, but it is especially important when merging files. Check to make sure that you end up with the number of observations that you expect. If you see a note like the following, you should be very suspicious. For this merge, there were 1,855 household observations, and 4,464 individual observations. I should end up with as many observations as there are people, and I did. I ended up with 4,464 cases, each with all of the household and person questions.

Note that the match command used the “TABLE” option. This is because it is a one-to-many match. In a one-to-many match, the file preceded with “Table” has one case per ID, and the file(s) preceded with file may have a varying number of cases per ID. The Household information was appended to each person in the household. Failure to specify this properly will result in an error message something like:

```
File #2
     KEY:        1        1

>Warning # 5132
>Duplicate key in a file.  The BY variables do not uniquely identify each
>case on the indicated file.  Please check the results carefully.
```

Your resulting data set will have valid household data for only the first person in each household.

If the file preceded with “Table” does NOT uniquely identify each case, you will get a message like:

```
>Error # 5131
>Duplicate key on a TABLE file.  Each case on a TABLE file in MATCH FILES
>must be uniquely identified by the BY variables.
>This command not executed.
```

Be sure to check your output for messages like this!

Now we will write a program to use the data set.

```
*** PROGRAM BEGINS HERE *** .
```

```
* vn95u2 sps .
title 'c:\all\SPSSclass\vn95\vn95u2.sps' .
* Documentation and data from following sites .
* http://csde.washington.edu/csde/vietnam/documents.html .
* http://csde.washington.edu/csde/vietnam/data.html .
** Read SPSS data set created by vn95.SPSS .
* Select only ever-married women, calculate age .
* Use "DO REPEAT"  .

get file = 'c:\all\SPSSclass\vn95\data\vn95.sav' .

* Create an ID number for all people in HH so that we can use it .
* for selection and identification .
compute idvar = 1 .
create idvar = csum(idvar) .
variable label idvar 'Identification Variable' .
execute .
* test variable - should range from 1 to 4464.
desc var = idvar.

************ Selection begins *****************  .
* Select only ever-married women who are heads of households  .
* or wives of heads of households  .

* Select only ever-married women  .
* women only  .
select if I5Q1B = 2  .
* ever married only  .
select if I5C15 = 1  .

* The respondent is not asked about her relationship to head  .
* in the individual questionnaire.  We must get that information from  .
* the household questionnaire, and then look at the line number of the  .
* respondent.  Info for 14 people in HH is given, much check all.  .
* Line numbers  .

* ln is for line number ;
* rl is for relationship to head 1 = Household Head, 2 = Wife of head  .
do repeat
  ln = H5P1Q1 H5P2Q1 H5P3Q1 H5P4Q1 H5P5Q1 H5P6Q1 H5P7Q1 H5P8Q1 H5P9Q1
       H5P10Q1 H5P11Q01 H5P12Q01 H5P13Q01 H5P14Q01 /
  rl = H5P1Q3 H5P2Q3 H5P3Q3 H5P4Q3 H5P5Q3 H5P6Q3 H5P7Q3 H5P8Q3 H5P9Q3
       H5P10Q3 H5P11Q03 H5P12Q03 H5P13Q03 H5P14Q03  .
* create a variable called target if head or relationship to head .
if rl = 1 or rl = 2  target = ln  .
end repeat .

* Select only cases where line number = target  .
* These should be household heads and wives of head  .
select if I5Q1A2 = target  .
* test .
freq var = target .
************ Selection ends *****************  .

************* calculating age begins ************ .
* month of birth = I5C1A  .
* year of birth  = I5C1B  .
* date of interview = I5Q8  .

* The dates of interviews are in the form of 291095  .
* We need to extract the year from it  .
compute yrint = (I5Q8 - (trunc(I5Q8/100)*100)) + 1900  .

* Set I5C1B to missing if value = 9999  .
if I5C1B = 9999   I5C1B = -99  .
missing values I5C1B (-99) .
```

```
compute age = yrint - I5C1B  .
variable label age = 'Age, from  yr of birth and year of interview'  .
************* calculating age ends ************ .
freq var = age .

***** calculate age at first marriage begins ***********  .
* I5C22 - What year did your first marriage begin - value = 98 if current marriage .
compute yr1stmr = I5C22  .
* I5C16 - is year current marriage began .
missing values I5C16 (9998) .
if yr1stmr = 98   yr1stmr = I5C16  .
compute age1stmr = yr1stmr - I5C1B  .
variable label
  yr1stmr  = 'Year of 1st marriage'
  age1stmr = 'Age of 1st marriage'  .
***** calculate age at first marriage ends   ***********  .

freq var = I5C1B I5C22 yr1stmr age1stmr .


**********;
* Homework for next week.  Create variables that calculate the number of children each .
* respondent has had, the age of the oldest child, .
* and the age of the woman at her first birth.
* Bring the program to class.  We will use these variables in class.  Also, look in the .
* documentation and find variables that you think would be interesting predictors .
* of the variables you create.  We will use multiple regression!  .

* Variables pertaining to children of respondent.
* Sex of child                          I5E4C1  - I5E4C14 .
* Month of birth                        I5E5AC1 - I5E5AC14.
* Year of birth                         I5E5BC1 - I5E5BC14 .
* Child wanted                          I5E6C1  - I5E6C14.
* If alive, place of residence          I5E7C1  - I5E7C14.
* If died, age at death                 I5E8C1  - I5E8C14.
* Enrolled, grade level                 I5E9AC1 - I5E9AC14.
* Not enrolled, level completed         I5E9BC1 - I5E9BC14.
* Not enrolled, current occupation      I5E10C1 - I5E10C14.
```