Combining Datasets

Goal

There are a number of reasons you may wish to combine LIS files. You may wish to combine files from different countries/years in order to run a regression utilizing country/year indicators. You may wish to use household information in a person-level analysis (or vice versa). You may want to include individuals or households from the shadow files if you are studying a specific subset of the population.

In this exercise, you will be asked to merge household-level and person-level variables from one dataset in the same file.

Activity

Use the data for Belgium 2000. First access the household-level dataset and calculate the mean of the number of household members. Then create a new dataset at the individual level containing ***casenum***, ***hweight***, ***d4***, ***d7***, ***d22***, ***ppnum***, ***pweight***, ***page***, ***psex***, ***pnwage*** (by merging the household and person-level datasets) and carry out the following tasks:

- create a household-level counter containing the number of persons in each household and compare it with the household-level variable ***d4***; calculate its mean over both individuals and over households;
- compare the mean individual wage of wage earners by region;
- compare the average age of individuals living in households that own their residence outright (i.e., without mortgage) to that of individuals living in households who still have a mortgage on their residence.

Guidelines

- In order to merge two datasets, they must necessarily contain at least one "merging variable", i.e. a variable which links the observations of one dataset to those of the other dataset. The variable must have the same name in the two datasets and will link observations with the same value in the two datasets (in case of several merging variables, it will link the observations with the same combination of values). To link household and person LIS datasets, such a variable is the household identifier (***casenum***), which exists in both datasets and takes the same values for the household at the household level and all the individuals belonging to that household in the person-level file.
- When merging datasets using a common merging variable, make sure that each of the files is sorted on that merging variable. The person and household files are already sorted on ***casenum***, but when you make certain selections or other data manipulations, you could make it a habit to always sort before merging, just to be on the safe side. In that case, you will need to save some temporary files. When saving temporary files, you can place them in a directory at LIS. In order that your filename differs from others saving files at the same time, save your files using your name. For example, Janet would save her temporary file by typing:

```
save outfile= "U:\janet\<filename.sav>
```

- When merging person-level information to household-level-information, you are combining data of two different units. Inform SPSS of different units being used in the following way : the file with the smaller units (persons) is referred to as file , whereas the file with the larger units (households) is called table. Reminder: you may need to sort on the "by"-variable before merging:

```
match files file = <p-file> /table = <h-file> /by casenum .
```

- In order to calculate household-level statistics from a person-level file, you can select on household heads only (***ppnum***=1), so that you are sure you include exactly one observation per household.

- To produce results for subsets you can (besides selecting) in stead use the command <split file>. This is a simple way to get results for instance per region. Make sure to sort the file before splitting on the variable by which you want to split up.

```
sort cases by <varlist>.
split file by <varlist> .
```

Program

```
title "** BASICS II - Exercise 4 **" .

get file = be00h
     / keep = casenum hweight d4 .
descriptives variables =  d4 .
weight by hweight.
title "weighted " .
descriptives variables =  d4 .
* merging P and H.
match files    file = be00p
                 / table = be00h
                 / keep = casenum hweight d4 d7 d22 ppnum pweight page psex pnwage
                 / by casenum .
title "unweighted (number of obs) " .
descriptives variables =  d4 .
temporary.
select if ppnum eq 1.
descriptives variables =  d4 .
sort cases by d7.
split file by d7.
if pnwage gt 0 posnwage = pnwage.
descriptives variables =posnwage .
sort cases by d22.
split file by d22.
descriptives variables =page .
split file off.
title "weighted " .
weight by pweight.
descriptives variables =  d4 .
temporary.
select if ppnum eq 1.
descriptives variables =  d4 .
sort cases by d7.
split file by d7.
if pnwage gt 0 posnwage = pnwage.
descriptives variables =posnwage .
sort cases by d22.
split file by d22.
```

```
descriptives variables =page .
```

Results

		Number of observations	Mean
d4		*2,697*	*2.44*
Counter of household members	For all observations	*6,935*	*3.19*
	For household heads only	*2,697*	*2.44*
pnwage	Flanders	*1,552*	*670,909*
	Brussels	*234*	*732,045*
	Wallonia	*826*	*648,581*
page	Owners with mortgage	*3,246*	*27.4*
	Owners without mortgage	*2,256*	*54.0*

Comments

- You will have noticed that in this exercise the merge worked perfectly, i.e., all observations of the merging file were uniquely linked to one observation in the using file. This is always the case with LIS household and individual level files from the same dataset because all individuals belong to at least one and no more than one household.
- When calculating descriptive statistics, you should always be careful to choose the unit (and hence the weight) that make most sense for the calculation: household-level statistics (such as household-level counters) should be calculated over households (using the household-level weight), while person-level statistics should be calculated over persons (using the person-level weight).