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 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 in Stata, 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 $mydata\janet, replace
```

- In order to retain person-level information in your merged file, start with your person-level sample and merge using the household sample. Be sure that you sort on the same variable before merging:

```
use <h-file>, clear
```

…code for creating household file information

```
sort casenum
save $mydata\<yourname>, replace
use <p-file>, clear
```

…code for creating person file information

```
sort casenum
merge casenum using $mydata\<yourname>
```

- When Stata does the merge, it automatically creates a variable called ***_merge,*** which flags cases where the merging variable(s) does not uniquely link observations across datasets (i.e. the value of the merging variable for one observation in dataset A does not exist in any of the observations of dataset B, in which case it takes the value 1, or vice versa, in which case it takes the value 2); in all cases where the observations uniquely match between the two datasets, the variable ***_merge*** will take the value 3. In order to retain only information that is present in both the household and person samples:

```
keep if _merge==3
drop _merge
```

- As of version 11, the merge command in Stata has been significantly altered. The same results as above can now be achieved in two lines:

```
use <varlistp> <p-file>, clear
merge m:1 casenum using <h-file>, keepusing(<varlisth>) keep(match)
nogen
```

- The **egen** command generates a new variable that assigns values to each observation over a user-defined group of variables. The **count** function of the **egen** command will calculate the number of valid (non-missing) observations of a chosen variable. To generate the number of observations within a household, define the group as the household id (***casenum***) and select a variable that is valid for all observations in the household as the counter variable. (The variable ***hweight*** will work well as a counter variable since it is never missing.) The final command will then look like:

```
egen tot  = count(hweight), by(casenum)
```

  The **egen** command is also useful for calculating the average age within a household:

```
egen meanage  = mean(page), by(casenum),
```

  the average wages within a household:

```
egen meanwage  = mean(pnwage), by(casenum);
```

  or the sum of all wages in a household:

```
egen sumwage  = sum(pnwage), by(casenum).
```

- In order to compare two variables in a dataset, you can use the **compare** command in the following way:

```
compare <var1> <var2>
```

- 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.

Program

```
di "** BASICS II - Exercise 4 (option 1)**"

use casenum hweight d4 d7 d22 using $be00h, clear
sum d4 [w=hweight]
sort casenum
save $mydata\teresa, replace

use casenum ppnum pweight page psex pnwage using $be00p, clear
sort casenum
merge casenum using $mydata\teresa
keep if _merge==3

di "Household level variables"
egen tot=count(casenum), by(casenum)
compare tot d4

sum tot [w=pweight]
sum tot if ppnum==1 [w=hweight]

di "Individual level variables"
bysort d7: sum pnwage if pnwage>0 [w=pweight]
bysort d22: sum page [w=pweight]

di "** BASICS II - Exercise 4 (option 2)**"
use casenum ppnum pweight page psex pnwage using $be00p, clear
merge m:1 casenum using $be00h, keepusing(casenum hweight d4 d7 d22) keep(match)
nogen

di "Household level variables"
egen tot=count(casenum), by(casenum)
compare tot d4

sum tot [w=pweight]
sum tot if ppnum==1 [w=hweight]

di "Individual level variables"
```

```
bysort d7: sum pnwage if pnwage>0 [w=pweight]
bysort d22: sum page [w=pweight]
```

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).