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-level 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 SAS, you will need to first sort your datasets by the variables used to link the datasets (***casenum).*** In order to retain person-level information in your merged file, use the **MERGE** statement following by a **BY** statement. Be sure that you sort datasets on the same variable before merging:

```
DATA mergefile;
  MERGE householdfile Individualfile;
     BY casenum;
…
```

```
RUN;
```

- When SAS sorts a dataset, it automatically creates **FIRST.** And **LAST.** Variables for each variable named in the **BY** statement. The value of ***FIRST.variable*** is 1 for the first observation with a given **BY** value and 0 for other observations. Similarly, the value of ***LAST.variable*** is 1 for the last observation for a given BY value and 0 for other observations. By default, SAS assumes that data being read with a **BY** statement are in ascending order of the **BY** values.
  - In order to create a household-level counter containing the number of persons in each household, you can use the following code. For each first observation of a household, the counter (comp) is initialised to 0 and then incremented by 1 (comp + 1) for each of the next member of the considered household.

```
IF FIRST.casenum THEN comp = 0 ;
   comp + 1 ;
```

  - Note that SAS treats differently `comp=comp+1;` and `comp+1;` In the first case, SAS add 1 to the value of the read observation whereas in the second piece of code, SAS increment by 1 the value of the preceding observation.

- In order to compare two variables in a dataset, you can use the **PROC COMPARE** procedure in the following way. Ensure that ***var1*** belongs to ***dataset1*** and ***var2*** belongs to ***dataset2***

```
PROC COMPARE BASE=dataset1 COMPARE=dataset2;
  VAR  var1 ;
  WITH var2 ;
RUN;
```

- 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

```
OPTIONS NOSOURCE NONOTES NOFMTERR NODATE NOCENTER LABEL NONUMBER LS=MAX
PS=MAX ;

DATA combh ;
 SET &be00h (KEEP=casenum hweight d4 d7 d22);
RUN;
PROC MEANS DATA=combh N MEAN;
  VAR d4 ;
  WEIGHT hweight ;
RUN;
PROC SORT DATA=combh ;
  BY casenum ;
RUN;
DATA combp ;
 SET &be00p (KEEP=casenum ppnum pweight page psex pnwage);
RUN ;
PROC SORT DATA=combp ;
  BY casenum;
RUN;
DATA test (KEEP=casenum comp) ;
 SET combp ;
  BY casenum ;
  IF FIRST.casenum THEN comp = 0 ;
     comp + 1 ;
IF LAST.casenum THEN  OUTPUT   ;
RUN;
PROC COMPARE BASE=combh COMPARE=test ;
  VAR  d4 ;
  WITH comp ;
RUN;
DATA comb ;
  MERGE combh combp ;
     BY casenum ;
RUN;
PROC MEANS DATA=comb N MEAN ;
  WHERE ppnum = 1 ;
  VAR d4 ;
  WEIGHT pweight ;
RUN;
PROC SORT DATA=comb ;
  BY d7 ;
RUN;
PROC MEANS DATA=comb N MEAN ;
  WHERE pnwage > 0 ;
  VAR pnwage ;
  WEIGHT pweight ;
  BY  d7 ;
RUN;
PROC SORT DATA=comb ;
  BY d22 ;
RUN;
PROC MEANS DATA=comb N MEAN ;
  WHERE d22 in (4,5) ;
```

```
  VAR page ;
  WEIGHT pweight ;
  BY  d22 ;
RUN;
```

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).