A Cross-national Comparison Using Stacked Data

Goal

In this exercise, we combine household- and person-level files across countries to run a regression estimating the usual hours of the working-aged civilian population, utilizing a number of the techniques learned in earlier exercises. In addition, we show the importance of weighting correctly when combining files with very diverse sample sizes and weighting structures.

Activity

For Ireland and Austria in 2000, do the following:

1. Using the household-level file:
    a. Create a sample containing ***casenum***, ***hweight***, ***d4***, ***d5***, and ***dpi***.
    b. Create a dummy variable (***poorhh***) indicating that a household is poor (as defined by having less than 50 percent of median equivalised disposable income).
    c. Create a variable (***count***) containing the number of households in the sample (that is the same for all observations within the country).
    d. Create a new weight that normalizes the household weight to 1.
    e. Save your file in the temporary LIS directory for later use. (See Guidelines below for details.)
2. Using the person-level file, create a sample of non-military working-aged adults (25 through 60).
    a. Include the variables ***country***, ***casenum***, ***pweight***, ***page***, ***psex***, ***peduc***, ***ptocc***, ***phoursu***, ***pclfs***, and ***pcare***.
    b. Using ***pclfs***, drop any individuals who are in the military (both regular armed forces and conscripts).
    c. Run the LIS *include* program to recode education. Recode ***educ*** to missing if education is not in one of the major classifications (1 to 3).
    d. Using ***pcare***, create a dummy variable (***ccare***) that indicates care of children.
    e. Recode ***phoursu*** to missing if the usual hours worked vary.
    f. Create dummy variables for gender (***female***) and country (***at00***).
3. For each country, merge the person- and household-level files keeping only working-age civilian adults in households without missing or zero dpi.
    a. Note: Your final file will be at the person level. You should only retain observations that include information from both of the two samples you created above.)
    b. Save the merged file with the same temporary file name as you used when saving the household data.
4. Create your final sample by appending all of your country-level merged files.
    a. Create a variable that defines each household in your combined sample (i.e., every combination of ***country*** and ***casenum*** will have a unique value).
    b. Save your final combined file to the temporary directory.

5. For this exercise, assume that the correct population model estimating the number of hours worked is as follows:

   ***phoursu = f(page educ poorhh female ccare* <country dummies>)**

   where <***country dummies***> are included for each country in your analysis (excluding, of course, the base country).

   a. Run a regression of your model for each country using the household weight. Since the sampling unit is the household, you will need to correct for the clustering of households.

   b. Run the same regression combining countries and including the dummy variable for Austria.

   c. Instead of using the household weight, use the normalized weight you created when you set up the household file running the regression using both countries.

Guidelines

- For household files, do not forget the standard data cleaning procedures (drop missing or 0 ***dpi***).
- When calculating equivalised household income, use the LIS equivalence scale (***dpi***/sqrt(***d4***)).
- For this exercise, no bottom- or top-coding is necessary.
- Stata will create temporary dummy variables for you if you use factor the `xi` command. Using `xi` will spare you from having to perform these extra steps and free up hard disk space. This is a very useful tool if you need to create dummies for categorical variable with a lot of values. In version 11, Stata has greatly expanded the interactive variables capabilities. See help for factor variables (***help fvvarlist***) for details.
- When saving temporary files, you can place them in a directory at LIS. In this exercise, you will create a file for each country. In order that your filename differs from others saving files at the same time, save your files using your name and add the four-digit country-year code for each new country. For example, Janet would save Austria by typing:

  ```
  save $mydata\janet_at00, replace
  ```

  and Ireland using:

  ```
  save $mydata\janet_ie00, replace
  ```

- Remember that the education routine to standardize education levels across countries can be called by:

  ```
  qui $myincl\educrecodepp.do
  ```

- The education recode program creates the variable ***educ*** that takes on the values of 1 (low) through 3 (high). (See Exercise 7 for more detail.) The value 9 is used for education levels that cannot be categorized.
- In some cases, surveys provide information that usual hours of work vary. Rather than coding these as missing and losing information, LIS codes variable usual hours in ***phoursu*** as -9. Users need to correct for this when estimating hours worked.
- In order to retain person-level information in your merged file, start with your person-level sample and merge using the household sample.

  ```
  use <varlistp> $at00p, clear
  merge m:1 casenum using $at00h, keepusing(<varlisth>) keep(match) nogen
  ```

  Then resave the file for later use:

  ```
  save $mydata\<yourname>_at00, replace
  ```

- To put all of your country files together, you will need to append the data sets:

```
use $mydata\<yourname>_ie00, clear
append using mydata\<yourname>_at00
```

  Then save the final file

```
save $mydata\<yourname>, replace
```

  Another way to combine data sets is to create a **tempfile** that disappears when the program ends:

```
tempfile temp
save `temp'
```

  If you need to save multiple files in your analyses, please investigate the use of the **tempfile** command.

  NOTE: Try to limit the number of files you save on the LISSY system. For this example, we saved three files in order to be clear about what we are doing. It is possible, however, to accomplish these same tasks without saving files.

- Instead of using the append command above, you can use the following loop if you want to append multiple files:

```
use $mydata\<yourname>_ie00, clear
foreach ccyy in at00 gr00 uk99 {
  display "`ccyy'"
  global ccyy "`ccyy'"
  qui append using $mydata\<yourname>_$ccyy
}
save $mydata\<yourname>, replace
```

- To run the regressions for this exercise, you can set a subprogram. The variables you need in this subroutine were already defined when you created the household and person files.
- When running a regression, always use the **robust** option. You will have made the correction for heteroskedasticity if it is necessary, and it will not affect your results if it is not.
- The weight you use should be based on the primary sampling unit. If households were randomly sampled, but you are looking at a sample of individuals, use the **cluster** option (in the example below it is called ***cgroup***).
- Your final subroutine should look something like:

```
program define doregress
   bysort country: reg phoursu page educ female ccare poorhh
   [w=hweight], robust cluster(cgroup)
   reg phoursu page educ female ccare poorhh AT [w=hweight], robust
   cluster(cgroup)
   reg phoursu page educ female ccare poorhh AT [w=normweight], robust
   cluster(cgroup)
end
```

Program

```
di "** STACKED DATA- Exercise 19 **"

global username "PUT YOUR LIS USERNAME HERE"
program define dohhold
  use casenum hweight d4 d5 dpi if (!mi(dpi) & !(dpi==0) & !(d5==3)) using
  ${${ccyy}h}, clear
  *Equivalised income
  gen ey=(dpi/sqrt(d4))
  *Poverty dummy
  qui sum ey [w=hweight], de
  gen byte poorhh=(ey<.5*r(p50))
  *Create a normalized weight
  egen normweight=sum(hweight)
  replace normweight=hweight/normweight
  *Create household counts
  egen count=count(casenum)
  sort casenum
  save $mydata\${username}_$ccyy, replace
end

program define doperson
  use country casenum page psex peduc ptocc phoursu pclfs pcare if
  inrange(page,25,60) using ${${ccyy}p} , clear
  drop if inrange(pclfs,180,189)
  qui do $myincl\educrecodepp.do
  recode educ 9=.
  recode pcare (100/119=1) (120/299 901=0) (-1 902=.), gen(ccare)
  recode phoursu (-9=.)
  recode psex (1=0) (2=1), gen(female)
  recode country (139=1) (137=0), gen(AT)
  sort casenum
end

program define domerge
  merge casenum using $mydata\${username}_$ccyy
  keep if _merge==3
  save $mydata\${username}_$ccyy, replace
end
```

```
program define doappend
  use $mydata\${username}_ie00, clear
  qui append using $mydata\${username}_at00
  *Define clusters consisting of each household unit
  egen cgroup=group(country casenum)
  save $mydata\${username}, replace
end

program define doregress
  bysort country: reg phoursu page educ female ccare poorhh [w=hweight], robust
   cluster(cgroup)
  reg phoursu page educ female ccare poorhh AT [w=hweight], robust
   cluster(cgroup)
  reg phoursu page educ female ccare poorhh AT [w=normweight], robust
   cluster(cgroup)
end

foreach ccyy in ie00 at00 {
  display "`ccyy'"
  global ccyy "`ccyy'"
  dohhold
  qui doperson
  qui domerge
}

qui doappend
    doregress
```

Results

		Weighted by ***hweight***			Normalised weight
		Ireland	Austria	Both	
Sum of weights		*$1.4*10^6$*	*$3.0*10^3$*	*$1.4*10^6$*	*2.5*
Coefficient / P>|t|	Age	*-.26*	*-.41*	*-.26*	*-.35*
		0.000	*0.000*	*0.000*	*0.000*
	Education	*2.66*	*3.38*	*2.66*	*2.75*
		0.000	*0.000*	*0.000*	*0.000*
	Female	*-19.64*	*-12.24*	*-19.62*	*-16.07*
		0.000	*0.000*	*0.000*	*0.000*
	Care for Child	*-2.91*	*-5.49*	*-2.91*	*-4.09*
		0.002	*0.000*	*0.002*	*0.000*
	Poor household	*-18.52*	*-2.97*	*-18.51*	*-13.18*
		0.000	*0.307*	*0.000*	*0.000*
	Austria	---	---	*1.51*	*1.80*
		---	---	*0.015*	*0.004*
	Constant	*47.25*	*49.95*	*47.26*	*48.65*
		0.000	*0.000*	*0.000*	*0.000*

Comments

- Weights in Ireland inflate to the population, while weights in Austria sum to 1 and inflate to the sample size. The results you see when running the two countries together are, therefore, being driven completely off of the results from Ireland.
- Be sure you understand the weighting procedures in each country. If you have one country with inflating weights and another without, it is vital that you normalize the weights across countries. If all countries in your analysis inflate to population weights, you can keep the weights as they are if you are interested in weighting all households equally, but you should normalize if you want to give each country equal importance. **Make sure you know what you are estimating!**