

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.
- As already mentioned in Exercise 5, 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 as suffix. For example, Janet would save Austria by typing:

```
save outfile = "U:\at_JC" .
```

and Ireland using:

```
save outfile = "U:\ie_JC" .
```

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


```
include file = "I:\educrcodepp.sav" .
```
- The education recode program creates the variable *educ* that takes on the values of 1 (low) through 3 (high). (See Exercise 8 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.
- Like in Exercise 5, merge person-level information to the household sample.

```
match files
  / file = "u:\atp_<your_initials>"
  / table = "u:\ath_<your_initials>"
  / by casenum .
```

Then save the combined file for later use:

```
save outfile "U:\att_<your_initials>" .
```

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

```
add files file = "u:/iet_pa"
  / file = "u:/att_pa".
```

- Your regression should look something like:

```
regression
  / regwgt = hweight
  / dependent = phoursu
  / method = enter page educ female ccare poorhh .
```

Program

```
title "*** STACKED DATA- Exercise 19 ***" .

get file = ie00p /keep = country casenum
      page psex peduc ptocc phoursu pclfs
      pcare.

select if page ge 25 and page le 60 .
select if pclfs le 179 or pclfs ge 190 .
include file = 'i:\educrcodepp.sps' .
if educ eq 9 educ = $sysmis .
if phoursu eq -9 phoursu = $sysmis .
recode pcare (100 thru 119 = 1)
      (120 thru 299 , 901 = 0) (-1 , 902 = sysmis) into ccare .
recode psex (1=0) (2=1) into female.
recode country (139 = 1) (137 = 0) into at .
save outfile = 'u:\iep_pa' .

get file = ie00h
      /keep = casenum hweight d4 d5 dpi.
select if dpi ne 0 .
select if not missing(dpi) .
compute group = 1 .
aggregate outfile = "u:\sumhwgt"
      / presorted
      / break = group
      / sumw = sum(hweight) .
match files
      / file = *
      / table = "u:\sumhwgt"
      / by group .
compute normw = hweight / sumw .
weight by hweight.
compute ey = dpi/(d4** 0.5) .
include file = 'i:\match-medianey.sps' .
compute povline = medianey * 0.5 .
compute poorhh = 0 .
if ey lt povline poorhh = 1 .
save outfile = 'u:\ieh_pa' .
```

```

match files
  / file = "u:\iep_pa"
  / table = "u:\ieh_pa"
  / by casenum .
descriptives hweight normw
  /statistics = default sum .
weight off.
regression
  / regwgt = hweight
  / dependent = phoursu
  / method = enter page educ female ccare poorhh .
save outfile = "u:/iet_pa" .

get file = at00p /keep = country casenum
      page psex peduc ptocc phoursu pclfs
      pcare.

select if page ge 25 and page le 60 .
select if pclfs le 179 or pclfs ge 190 .
include file = 'i:\educrcodepp.sps' .
if educ eq 9 educ = $sysmis .
if phoursu eq -9 phoursu = $sysmis .
recode pcare (100 thru 119 = 1)
      (120 thru 299 , 901 = 0) (-1 , 902 = sysmis) into ccare .
recode psex (1=0) (2=1) into female.
recode country (139 = 1) (137 = 0) into at .
save outfile = 'u:\atp_pa' .

get file = at00h
  /keep = casenum hweight d4 d5 dpi.
select if dpi ne 0 .
select if not missing(dpi) .
compute group = 1 .
aggregate outfile = "u:\sumhwgt"
  / presorted
  / break = group
  / sumw = sum(hweight) .
match files
  / file = *
  / table = "u:\sumhwgt"

```

```

    / by group .
compute normw = hweight / sumw .
weight by hweight.
compute ey = dpi/(d4** 0.5) .
include file = 'i:\match-medianey.sps' .
compute povline = medianey * 0.5 .
compute poorhh = 0 .
if ey lt povline poorhh = 1 .
save outfile = 'u:\ath_pa' .

match files
    / file = "u:\atp_pa"
    / table = "u:\ath_pa"
    / by casenum .
descriptives hweight normw
    /statistics = default sum .
weight off .
regression
    / regwgt = hweight
    / dependent = phoursu
    / method = enter page educ female ccare poorhh .
save outfile = "u:/att_pa"

add files file = "u:/iet_pa"
    / file = "u:/att_pa" .
descriptives hweight normw
    / statistics = mean sum .
weight off .
regression
    / regwgt = hweight
    / dependent = phoursu
    / method = enter page educ female ccare poorhh at .
regression
    / regwgt = normw
    / dependent = phoursu
    / method = enter page educ female ccare poorhh at .

```

Results

		Weighted by <i>hweight</i>			Normalised weight
		Ireland	Austria	Both	
Sum of weights		$1.4*10^6$	$3.0*10^3$	$1.4*10^6$	2.5
Coefficient <hr/> 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!**