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.

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

  ```
  %INCLUDE "i:\educrecodepp.sas";
  ```

- 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. Note that currently, you must recode ***educ*** to missing in a separate data step.

- 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.

- To put all of your country files together, you will need to append the data sets. Use the `PROC APPEND` procedure and use the option `FORCE` to avoid any SAS format errors generated if two different SAS formats are used for the same variables.

  ```
  PROC APPEND BASE=ie00 DATA=at00 FORCE;
  RUN;
  ```

- To create the ***cgroup*** variable (i) sort your dataset by ***country*** and ***casenum***, (ii) and create a counter that increment by 1 for each different household. Adapt the following code.

```
PROC SORT DATA=…;
 BY country casenum ;
 RUN;
DATA …;
 SET …;
  BY country casenum;
     RETAIN cgroup ;
     IF _N_ = 1 THEN cgroup = 0 ;
     IF FIRST.casenum THEN cgroup + 1;
RUN;
```

- To run the regressions for this exercise, you can set a subprogram using the `PROC SURVGEN` procedure.

  - 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` statement of the `PROC SURVGEN`.

  ```
  CLUSTER cgroup;
  ```

- Your final subroutine and the code to invoke it should look something like:

```
%MACRO regress;
  PROC SURVEYREG;
     &by;
     WEIGHT  &poids;
     MODEL   &modele;
     CLUSTER cgroup;
  RUN;
%MEND regress;

…

%LET by     = BY country;
%LET poids  = hweight   ;
%LET modele = phoursu = page educ poorhh female ccare   ;
%regress
```

Program

```
OPTIONS NOSOURCE NONOTES NOFMTERR NODATE NOCENTER LABEL NONUMBER LS=200
PS=MAX ;

%MACRO prep;

DATA h ;
 SET &&&pi.h (KEEP=casenum hweight d4 d5 dpi );
    IF dpi not in (. 0);
      IF d5 ne 3         ;
    ey = dpi /SQRT(d4) ;
RUN;
PROC UNIVARIATE DATA=h NOPRINT;
   VAR ey;
   WEIGHT hweight ;
   OUTPUT OUT=temp MEDIAN=medey;
RUN ;
DATA _NULL_;
 SET temp;
   CALL SYMPUT("m",medey);
RUN;
DATA h ;
  IF _N_ = 1 THEN
     DO UNTIL (eof) ;
        SET h END=eof;
          count + 1 ;
          sw + hweight ;
     END ;
 SET h ;
    normw = hweight / sw  ;
    poorhh = 0 ;
    IF ey < 0.5 * &m THEN poorhh = 1 ;
RUN ;
PROC SORT DATA=h ;
  BY casenum ;
RUN;

DATA p ;
 SET &&&pi.p (KEEP=country casenum page psex peduc ptocc phoursu pclfs
pcare);
    IF (25 <=page <= 60)             ;
    IF (180<=pclfs<=189) THEN DELETE ;
    IF      (100<=pcare<=119)                  THEN ccare = 1 ;
    ELSE IF ((120<=pcare<=299) OR (pcare=901)) THEN ccare = 0 ;
    ELSE                                            ccare = . ;
    IF phoursu = -9 THEN phoursu = . ;
    IF psex = 1 THEN female = 0 ;
    IF psex = 2 THEN female = 1 ;
    IF  country = 139 THEN at = 1 ;
    ELSE                   at = 0 ;
    %INCLUDE "i:\educrecodepp.sas";
RUN;
DATA p ;
 SET p ;
   IF educ = 9 THEN educ = . ;
```

```
RUN;
PROC SORT DATA=p ;
  BY casenum ;
RUN;
DATA &pi ;
 MERGE p h ;
    BY casenum ;
    IF country ne . ;
RUN;

%MEND prep ;

%MACRO regress ;
  PROC SURVEYREG;
     &by ;
     WEIGHT  &poids;
     MODEL   &modele;
     CLUSTER cgroup;
  RUN;
%MEND regress;

%LET pi = ie00;
%prep
%LET pi = at00;
%prep
PROC APPEND BASE=ie00 DATA=at00 FORCE;
RUN;
PROC SORT DATA=ie00;
 BY country casenum;
 RUN;
DATA surv ;
 SET ie00 ;
  BY country casenum;
     RETAIN cgroup ;
     IF _N_ = 1 THEN cgroup = 0;

     IF FIRST.casenum THEN cgroup + 1;
RUN;
%LET by     = BY country;
%LET poids  = hweight   ;
%LET modele = phoursu = page educ poorhh female ccare;
%regress
%LET by     = ;
%LET poids  = hweight   ;
%LET modele = phoursu = page educ poorhh female ccare at;
%regress
%LET by     = ;
%LET poids  = normw ;
%LET modele = phoursu = page educ poorhh female ccare at;
%regress
```

Results

<table>
<tr><td colspan="2" rowspan="2"></td><td colspan="3">Weighted by <b><i>hweight</i></b></td><td rowspan="2">Normalised weight</td></tr>
<tr><td>Ireland</td><td>Austria</td><td>Both</td></tr>
<tr><td colspan="2">Sum of weights</td><td>$1.4*10^{6}$</td><td>$3.0*10^{3}$</td><td>$1.4*10^{6}$</td><td>2.5</td></tr>
<tr><td rowspan="14">Coefficient
P>|t|</td><td rowspan="2">Age</td><td>-.26</td><td>-.41</td><td>-.26</td><td>-.35</td></tr>
<tr><td>0.000</td><td>0.000</td><td>0.000</td><td>0.000</td></tr>
<tr><td rowspan="2">Education</td><td>2.66</td><td>3.38</td><td>2.66</td><td>2.75</td></tr>
<tr><td>0.000</td><td>0.000</td><td>0.000</td><td>0.000</td></tr>
<tr><td rowspan="2">Female</td><td>-19.64</td><td>-12.24</td><td>-19.62</td><td>-16.07</td></tr>
<tr><td>0.000</td><td>0.000</td><td>0.000</td><td>0.000</td></tr>
<tr><td rowspan="2">Care for Child</td><td>-2.91</td><td>-5.49</td><td>-2.91</td><td>-4.09</td></tr>
<tr><td>0.002</td><td>0.000</td><td>0.002</td><td>0.000</td></tr>
<tr><td rowspan="2">Poor household</td><td>-18.52</td><td>-2.97</td><td>-18.51</td><td>-13.18</td></tr>
<tr><td>0.000</td><td>0.307</td><td>0.000</td><td>0.000</td></tr>
<tr><td rowspan="2">Austria</td><td>---</td><td>---</td><td>1.51</td><td>1.80</td></tr>
<tr><td>---</td><td>---</td><td>0.015</td><td>0.004</td></tr>
<tr><td rowspan="2">Constant</td><td>47.25</td><td>49.95</td><td>47.26</td><td>48.65</td></tr>
<tr><td>0.000</td><td>0.000</td><td>0.000</td><td>0.000</td></tr>
</table>

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!**