Nonresponse and dynamic models

Cheti Nicoletti
University of Essex

Sampling and nonsampling errors

- A first problem that may affect a survey is the under or over-representation of some groups of persons in the population.
- This problem may be due to two categories of errors: ***the sampling and the non sampling errors***.
- In the ECHP ***weights*** are computed to take account of the sampling design and household nonresponse (household fully nonrespondent).

Household nonresponse

- A household is responding if at least one eligible member returns the personal questionnaire and the household questionnaire, it is nonresponding otherwise.
- ECHP take account of nonresponding households by computing weights for all individuals belonging to responding households

Unit and item nonresponse for responding households

- *Unit nonresponse*: It occurs when an eligible individual fails to return the personal questionnaire;
- *Item nonresponse*: An individual who returns the questionnaire but does not respond to a specific question is said to be item nonresponding to that question (ex. income questions).

Weighting for the sampling design and for individuals in responding households

- ***Starting weights*** are design weights computed to take account of the different sample selection probabilities in the first wave. (starting weights are equal to the base weights in previous wave for wave>1)
- Starting weights are multiplied by a factor to take account of the probability for an individual being resident in an interviewed household.

Computation of base and cross sectional weights

- ***Base weights*** for *sample persons* are then computed by applying a calibration (post-weighting) to reflect the population structure by age and sex and to reflect the marginal distributions for the variables household size, tenure status, number of economically active persons and regions.
- ***Cross sectional weights*** for *sample and nonsample persons* are computed as the average of the base weights of the household members
- ***Household cross-sectional weights*** are equal to the personal cross-sectional weights of the household members.

Weighting and imputing for unit nonresponse

- When we use information from the personal questionnaires, we must also take account of unit nonresponses (personal interview not completed at all) in responding households.
- For income variables this type of non response problem is solved by applying an imputation procedure.
- For all other variables ECHP allows to take account of the unit nonresponse in a responding household using adequate weights: the weights for interviewed persons.

Weights for interviewed persons

- ***Base weights for interviewed persons*** are equal to the base weights for sample persons divided by the probability of being interviewed, when eligible.
- ***Cross-sectional weights for interviewed persons*** are equal to the average base weights for interviewed persons within households.

List of weights available in the ECHP

- RG003 BASE WEIGHTS are weights for sample persons in responding households (0 for non sample persons)
- RG002 PERSONAL WEIGHTS (CROSS-SECTIONAL) are weights for sample and non sample persons resident in responding households (computed as the average rg003 within households)
- HG004 HOUSEHOLD CROSS-SECTIONAL WEIGHTS are equal to the personal cross-sectional weights, rg002, of the household members.
- PG003 BASE WEIGHTS are weights for interviewed sample persons.
- PG002 PERSONAL WEIGHTS (CROSS-SECTIONAL) are weights for interviewed sample and non sample persons (computed as the average pg003 within households).

How to use the weights

- RG003 BASE WEIGHTS should be used for inference on sample individuals and for variables from the register files which are available for all members in responding households.
- RG002 PERSONAL WEIGHTS (CROSS-SECTIONAL) should be used for inference on sample and nonsample individuals and for variables from the register files which are available for all members in responding households
- HG004 HOUSEHOLD CROSS-SECTIONAL WEIGHTS should be used for inference on responding households and for variables in the household files.
- PG003 BASE WEIGHTS should be used for inference on interviewed sample individuals and for variables from the personal files
- PG002 PERSONAL WEIGHTS (CROSS-SECTIONAL) should be used for inference on interviewed sample and nonsample individuals and for variables from the personal files.

What is the aim of the weights?

- ECHP computes weights to take account of different sampling probabilities and of the household and unit nonresponses, while it takes account of unit and item non-response for income variables using imputation procedures.

References:

1. Eurostat (2003), ``Construction of wieghts in the ECHP'', PAN 165.
2. Eurostat (2003), ``Imputation of income in the ECHP'', PAN 164.

How to use weights in Stata

- Most Stata commands can deal with weighted data. Stata allows four kinds of weights:
  1. fweights, or frequency weights, are weights that indicate the number of duplicated observations.
  2. ***pweights, or sampling weights, are weights that denote the inverse of the probability that the observation is included due to the sampling design.***
  3. aweights, or analytic weights, are weights that are inversely proportional to the variance of an observation; i.e., the variance of the j-th observation is assumed to be sigma^2/w_j, where w_j are the weights.
  4. iweights, or importance weights, are weights that indicate the "importance" of the observation in some vague sense.

Weighted regressions

- If we use only interviewed sample persons

regress y x1 x2 x3 [pweight=pg003]

- If we use interviewed sample and nonsample persons

regress y x1 x2 x3 [pweight=pg002]

- If we use all sample persons in responding households

regress y x1 x2 x3 [pweight=rg003]

- If we use all sample and nonsample persons in responding households

regress y x1 x2 x3 [pweight=rg002]

- If we use all responding households

regress y x1 x2 x3 [pweight=hg004]

Missing personal income components are computed by an imputation procedure at the individual level called ***Imputation and Variance Estimation, IVE*** .

- The IVE procedure is repeated iteratively until the difference between the imputed values obtained from two consecutive iterations is lower than a given threshold or the number of iterations exceeds a given number.
- The imputation procedure proceeds by steps.
  - First imputation is applied to variables with a low fraction of missing cases using information on variables without missing data.
  - Then imputation is applied to variables with more severe missing problem, conditioning both on variables without missing data and variables imputed in the first step;
  - and so on.
- The specific model used for the imputation depends on the type of variable to be imputed. For example, it is a linear regression model when the target variable is continuous and a logistic regression model when the target variable is binary.

Imputation and Variance Estimation (IVE) procedure.

- In the initial stage, the auxiliary variables used for imputation are sex, age, employment characteristics (socio-professional category, employment sector, size of the firm, type of job, hours worked per week and education level). Even these variables are sometimes missing, and so they become target variables to be imputed at a previous step of the IVE procedure.
- For the imputation of a specific target variable past information may also be used. In particular, the value observed for the target variable in the previous wave is used as an auxiliary variable for the imputation of its current value, but not for the imputation of other variables. If the value of the target variable in last wave is not observed but imputed, it is not used.
- The IVE procedure allows to define a range for the variable to be imputed. In the ECHP this range is equal to the observed range for responding people, that is imputed value must lie between the minimum and the maximum values observed for the responding persons.

References for the IVE procedure

- The imputation is performed using the software *IVEware*, which is also used for imputation of variables in the Panel Study on Income Dynamics in USA.
- Raghunathan T.E., Solenberger P.W., Hoewyk J.V. (1999), "*IVEware: Imputation and Variance Estimation Software. Installation Instructions and User Guide. Survey Methodology Program*" Survey Research Center, Institute for Social Research, University of Michigan.

Definition of income variables in the ECHP

The income variables consist of annual amounts in the year before the survey, net of taxes and expressed in national units and current prices.

Exceptions

- France and Finland collect and report all income components as gross.
- The ECHP provides also information of monthly earning in the month before the survey.

The income components may be classified

By type of income sources

- wages and salaries,
- income from self-employment or farming,
- pensions (old-age related benefits and survivors' benefits),
- unemployment/redundancy benefits,
- other social benefits or grants (family-related allowances, sickness/invalidity benefits, education-related allowances, other personal benefits, social assistance, housing allowances),
- nonwork private income (capital income, property/rental income, private transfers received).

Imputing household income when item nonresponse occurs

$$Y_h^I = \sum_{i,j} D_{hi}\left[R_{hij}Y_{hij} + (1-R_{hij})\hat{Y}_{hij}\right]$$

$D_{hi} = \text{indicator of unit response}$

$R_{hij} = \text{indicator of item response on income component j}$

$\hat{Y}_{hij} = \text{imputed personal income for component j}$

$Y_{hii} = \text{reported income component j}$

$Y_h^I = \text{imputed household income}$

Imputing household income when unit nonresponse occurs in the ECHP-UDB 2004

- **For households with unit nonresponse, household income is obtained according to the following formula**

$$Y_h^F = Y_h^I + \sum_i (1 - D_{hi}) Y_{hi}^A$$

$Y_h^{\mathrm{I}} = \text{imputed household income}$

$D_{hi} = \text{indicator of unit response}$

$Y_{hi}^A = \text{additional income for unit nonrespondents}$

Imputation for unit nonresponse

- The imputation for unit nonresponse uses as auxiliary variables personal and household characteristics from the current and the previous waves.
- When a household includes unit nonrespondents, total household income is computed by summing up reported and imputed personal incomes. The resulting amount is then compared with that obtained multiplying by 12 the total monthly household income reported in the last wave. The additional household income is equal to the difference between the former and the latter amount if this difference is positive, and to zero otherwise.
- If the household composition changes between waves or the monthly household income is missing in the last wave, then the current monthly household income is used instead. For the first wave of the panel, information on past income is not available and the values from the following waves are used.

Imputing household income when unit nonresponse occurs in the ECHP-UDB 2002

- No lagged and individual variables are used to compute

$$Y_h^F = f_h Y_h^I = f_h \sum_i D_{hi} \left[R_{hi} Y_{hi} + (1 - R_{hi}) \hat{Y}_{hi} \right]$$

$$f_h = \text{within - household nonresponse infation factor}$$

Information on imputation available in the ECHP

- No information is available at the individual level.
- Two indices are available at the household level:
  1. imputation index for unit nonresponse within responding households,
  2. imputation index for the item nonresponse

Total net household income

is obtained by summing up over different types of income and over the individuals belonging to the same household.

We distinguish between the following different types of nonresponse on household income:

1. ***Full item nonresponse***: It occurs when all income components are missing.
2. ***Partial item nonresponse***: It occurs when only some income components are missing. This latter category contains households with:
    i. ***only item nonresponse***,
    ii. ***only unit nonresponse***, and
    iii. ***both unit and item nonresponse***.

We can compute 3 imputation indices at household level

$\frac{Y_h^R}{Y_h^I}$ item imputation index

$\frac{Y_h^I}{Y_h^F}$ unit imputation index

$\frac{Y_h^R}{Y_h^F} = \frac{Y_h^R}{Y_h^I}\frac{Y_h^I}{Y_h^F}$ total imputation index

The impact of the nonresponse and imputation on household income

- We focus attention on the total net household income (equivalized) for responding households.
- The aim is the evaluation of the imputation procedures adopted in the ECHP by comparing the structure of the of the household income for households with
  1. complete response,
  2. only item nonresponse,
  3. full item nonresponse,
  4. only unit nonresponse and
  5. both item and unit nonresponse.

ANALYSIS OF THE EQUIVALIZED HOUSEHOLD INCOME

- **TYPES OF ANALYSIS:**

- Comparison of mean and median of the equivalized household income for different types of nonresponse
- Median regressions to control for a set of explanatory variables (sex, cohabitation, education, age, household size and number of children) and separately for different types of responding categories.

- **MAIN RESULTS:**

- It seems that the average and median household income (conditional or marginal) is lower for full item nonresponding households and higher for households with both unit and item nonresponses relatively to fully responding households.
- In the ECHP-UDB 2002 a quite big fall in the pseudo R2 for the median regression is observed for households with full income nonresponse, with unit nonresponse and with both unit and item nonresponse. The fall in the pseudo R2 is observed instead only for the households full income nonrespondents when using the more recent ECHP-UDB 2003.

Table 1: Average values of imputation ratios and of household income by type of nonresponse.

	Complete response	Type of nonresponse				Total
		Only item	Full item	Only unit	Item/Unit	
2002 UDB						
Item imputation ratio	1	0.791	0.000	1.000	0.636	0.934
Unit imputation ratio	1	1.000	1.000	697	0.662	0.985
Total imputation ratio	1	0.791	0.000	0.697	0.457	0.926
Reported income	9.175	8.495	0	6.121	4.684	8.720
Item imputed income	9.175	10.763	6.266	6.120	7.056	9.293
Final income	9.176	10.763	6.266	8.985	11.225	9.476

Table 1: Average values of imputation ratios and of household income by type of nonresponse.

	Complete response	Type of nonresponse				Total
		Only item	Full item	Only unit	Item/Unit	
2002 UDB						
Median regression						
Intercept	9.297	10.315	5.182	8.585	10.552	9.475
Standard error	0.012	0.034	0.081	0.572	0.053	0.012
Adjusted R^2	.29	.26	.08	.10	.21	
No. obs.	562,685	118,452	11,652	678	31,229	724,696
Percentage	77.6	16.3	1.6	0.1	4.3	100

Table 1: Average values of imputation ratios and of household income by type of nonresponse.

	Complete response	Type of nonresponse				Total
		Only item	Full item	Only unit	Item/Unit	
2003 UDB						
Item imputation index	1	0.798	0	1	0.605	0.936
Unit imputation index	1	1	1	0.677	0.624	0.988
Total imputation index	1	0.798	0	0.677	0.434	0.931
Reported income	9.137	8.348	0	6.287	4.255	8.700
Item imputed income	9.137	10.344	6.339	6.287	6.195	9.208
Final income	9.137	10.345	6.339	9.943	10.095	9.329
Median regression						
Intercept	9.336	10.282	5.364	10.043	10.078	9.483
Standard error	0.013	0.025	0.063	0.050	0.073	0.011
Adjusted R^2	.29	.26	.07	.30	.31	
No. obs.	572,089	126,649	12,050	517	22,723	734,028
Percentage	77.9	17.3	1.6	.1	3.1	100

Problems in dynamic models

1. Dynamic models for continuous dependent variables
   - When considering first differences to control for unobserved heterogeneity there is a *correlation between error and lagged dependent variable*
2. Dynamic models for discrete or categorical dependent variables
   - *Initial condition problem*
   - *State dependence versus heterogeneity*

Dynamic regression model

$y_{i,t} = \rho\, y_{i,t-1} + x_{i,t}\,\beta + \mu_i + u_{i,t}$ *level*

$dy_{i,t} = \rho\, dy_{i,t-1} + dx_{i,t}\,\beta + du_{i,t}$ *first differences*

- Fixed effects estimator applied to dynamic models is inconsistent because $dy_{i,t-1}$ *and* $du_{i,t}$ *are correlated*
- The solution is to use lagged *y* as IV and apply a GMM estimator.
- Arrelano-Bond linear, dynamic panel-data estimation
- Stata command: xtabond

Initial condition problem

(1) $Y^*_{i,t} = \rho\, y_{i,t-1} + x_{i,t}\,\beta + \mu_i + u_{i,t}$ t=2,…T where $y_{i,t} = I(Y^*_{i,t} > 0)$

(2) $Y^*_{i,1} = z_{i,}\,\gamma + \eta_i + e_{i,t}$ for t=1 (we do not know $y_{i,0}$)

If η_i and μ_i are correlated then $y_{i,1}$ is correlated with μ_i in

$$Y^*_{i,2} = \rho\, y_{i,1} + x_{i,2}\,\beta + \mu_i + u_{i,2}$$

and we cannot estimate consistently ρ and β

Solution: joint estimation of equation (1) and initial condition (2) allowing correlation between μ_i and $\eta_{i.}$

State dependence versus heterogeneity

$\Pr(y_{i,t}=1|y_{i,t-1}=0,x_{i,t})\neq\Pr(y_{i,t}=1|y_{i,t-1}=1,x_{i,t})$

might imply state dependence

- If $\Pr(y_{i,t}=1|y_{i,t-1}=0,x_{i,t})\neq\Pr(y_{i,t}=1|y_{i,t-1}=1,x_{i,t})$ but $\Pr(y_{i,t}=1|y_{i,t-1}=0,x_{i,t},\mu_i)=\Pr(y_{i,t}=1|y_{i,t-1}=1,x_{i,t},\mu_i)$ then state dependence is spurious and due to unobserved heterogeneity.
- If $\Pr(y_{i,t}=1|y_{i,t-1}=0,x_{i,t})\neq\Pr(y_{i,t}=1|y_{i,t-1}=1,x_{i,t})$, $\Pr(y_{i,t}=1|y_{i,t-1}=0,x_{i,t},\mu_i)\neq\Pr(y_{i,t}=1|y_{i,t-1}=1,x_{i,t},\mu_i)$ and the error terms are uncorrelated then state dependence is true.

Chamberlain (1978) suggests (to control for the possible correlation in the errors)

$Y^*_{i,t} = \rho\,(x_{i,t-1}\,\beta + \mu_i + u_{i,t-1}) + x_{i,t}\,\beta + \mu_i + u_{i,t}$

- If $\Pr(y_{i,t}=1 | y_{i,t-1}=0, x_{i,t}) \neq \Pr(y_{i,t}=1 | y_{i,t-1}=1, x_{i,t})$, $\Pr(y_{i,t}=1 | y_{i,t-1}=0, x_{i,t}, \mu_i) \neq \Pr(y_{i,t}=1 | y_{i,t-1}=1, x_{i,t}, \mu_i)$ but $\Pr(y_{i,t}=1 | y_{i,t-1}=0, x_{i,t}, x_{i,t-1}, \ldots, \mu_i) = \Pr(y_{i,t}=1 | y_{i,t-1}=1, x_{i,t}, \mu_i)$ then there is no state dependence
- If $\Pr(y_{i,t}=1 | y_{i,t-1}=0, x_{i,t}) \neq \Pr(y_{i,t}=1 | y_{i,t-1}=1, x_{i,t})$, $\Pr(y_{i,t}=1 | y_{i,t-1}=0, x_{i,t}, \mu_i) \neq \Pr(y_{i,t}=1 | y_{i,t-1}=1, x_{i,t}, \mu_i)$ but $\Pr(y_{i,t}=1 | y_{i,t-1}=0, x_{i,t}, x_{i,t-1}, \ldots, \mu_i) \neq \Pr(y_{i,t}=1 | y_{i,t-1}=1, x_{i,t}, \mu_i)$ then there is state dependence

Duration models with the ECHP

- It is difficult to estimate duration models because:

1. the panel is not very long (8 years)
2. there are very few retrospective questions.

- Nevertheless, it is possible to estimate duration models by selecting a sample of people entering the state (whose duration we want to study) in first wave and then following them until the change state.

Stock and flow sample

- Flow sample: sample of individuals who enter a specific state at some time during an interval (ex: people entering unemployment during 1994).
- Stock sample: sample of individual who are observed in a specific state in a specific time point (ex: people unemployed when interviewed in 1994).

Stock sample gives a bias estimation of the average duration

- In the stock sample people with longer durations (unemployment spells) have higher probability to be selected.
- This implies an overestimation of the average duration (unemployment duration) when using stock sample instead then a flow sample.

Computing average household size and income using weights

```
 use country hid hd005 wave hi100 hg004 using
     "y:\all\stata\trn_w1h.dta", clear
 keep if country==8 | country==51|country==57
*dividing household income by the EQUIVALISED SIZE
gen ehincome=hi100/hd005
rename hi100 hincome
rename hd005 hsize
sort country
merge country using country
keep if _m==3
replace hincome=hincome/ppp1993
hincome was long now double
replace ehincome=ehincome/ppp1993
gen hincomew=hincome*hg004
gen ehincomew=ehincome*hg004
gen hsizew=hsize*hg004
```

The households with higher size are overrepresented in the sample

- Therefore the weights for households with smaller size are on average higher then the weights for bigger households.

```
 sum hg004 if hsize>2

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
       hg004 |      4811    .8090473    .6200689          0   11.38646

sum hg004 if hsize<=2

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
       hg004 |     10570    1.086913    .9204035          0   28.57697
```

The average household size decreases after using the weights

bys country: sum hincome hincomew ehincome ehincomew hsizew hsize

-> country = ireland

Variable	Obs	Mean	Std. Dev.	Min	Max
hincome	4038	21914.9	20721.32	90.42335	704386.9
hincomew	4038	19107.49	17651.49	98.97591	435945.1
ehincome	4036	10464.82	11534.7	60.28223	565361
ehincomew	4036	9845.492	10873.94	65.98394	415856.9
hsizew	4046	1.973523	1.240311	.21816	15.06672
hsize	4046	2.140064	.8237746	1	6.3

Balanced and unbalanced panels

```
use pfile1, clear
local i=2
while `i'<=8{
append using pfile`i'
local i=`i'+1
}
Save unbalanced, replace
sort country pid wave
by country pid: gen N=_N
tab N
keep if N==8
save balanced, replace
```

Panel with monotone attrition

```
use pfile1, clear
sort country pid
local i=2
while `i'<=8{
merge country pid using pfile`i'
tab _m
gen r`i'=(_m==3)
drop if _m==2
drop _m
sort country pid
local i=`i'+1
}
keep pid country r*
sort country pid
merge country pid using unbalanced
tab _m
keep if _m==3
drop _m
local i=2
while `i'<=8{
drop if wave>=`i' & r`i'==0
local i=`i'+1
}
save panelattr, replace
```