Initial Wave Nonresponse and Panel Attrition in the Finnish Subsample of EU-SILC

Tara Junes[1]

[1]Statistics Finland, University of Helsinki, e-mail: tara.junes@stat.fi

Abstract

The objective of this paper is to study the effects of initial wave unit nonresponse and panel attrition on the quintile distribution of disposable household equivalised income. Analyses are performed for one rotation group selected from Finnish EU-SILC. In addition to empirical analysis the changes between quintiles states are modelled with Markov chains.

Keywords: Unit nonresponse, Attrition, EU-SILC, Markov chains

1 Introduction

This paper is a summary of my Master's Thesis (Junes 2012). The purposes of the study was to investigate unit nonresponse in the Finnish subsample of EU Statistics of Income and Living Conditions (EU-SILC) which is a panel study with a four-year rotational sampling design.

The common perception is that estimation results of a panel study become more biased with increasing amount of attrition. However, it has been shown that a nonresponse bias at the beginning of the panel can fade away in subsequent panel waves without any correction. The fade-away phenomenon occurred in the analyses of certain income variables, from which the most interesting is disposable household equivalised income. (Rendtel et al. 2004 and Gerks 2004).

The fade-away theory was suggested by analysing the Finnish subsample of European Community Household Panel (ECHP). The objective of my Thesis was to investigate the fade-away hypothesis for a different dataset and to show that the existence of the hypothesis is not so straightforward. The dataset being analysed consists of information collected for one rotation group from Finnish EU-SILC. The selected rotation group can be seen as a non-rotational panel with duration of four years the initial wave being the first analyse year.

The main attention is given to computation of the income quintiles and to the empirical and theoretical modelling of transitions between computed quintiles. The research is performed for three different groups of sampling units:

1 All sampling units that were intended to participate the panel at the initial wave i.e. in 2005 (FULL-sample).
2 All respondents of the initial wave (RESP-sample).
3 All observed panel members in subsequent waves (OBS-sample).

The effects of initial wave nonresponse are displayed by comparing the income distribution of FULL-sample with the corresponding results from the RESP-sample. If there is differences between the RESP-sample and the OBS-sample, it is a possible sign of attrition bias.

The main analysis variable is disposable household equivalised income which is the total gross household income minus current transfers paid adjusted by the household composition. The adjustment is done with the OECD-modified equivalence scale assigning a value of 1 to the first household member, of 0.5

to each additional adult and of 0.3 to each child (Atkinson *et al.* 2002). The total disposable household income of a dwelling unit is divided by the sum of the scaling values after which the quotient is assigned to each individual in the household. In the previous analyses fade-away effect was seen in the distributions of both equivalised and non-equivalised household income (Gerks 2004).

2 Data

The Finnish EU-SILC sample 2006 was drawn from the Population Information System (PIS) maintained by the Population Register Centre of Finland. The number of persons belonging to the sample for the selected rotation group is 2 500. The rotating structure of the panel is displayed in Table 1. The income information for the sampling units was collected from the registers of Statistics Finland.

Table 1: Rotation structure of EU-SILC

	Measurement year								
Rotation group	1	2	3	4	5	6	7	8	...
R1	X	X	X	X					
R2		X	X	X	X				
R3			X	X	X	X			
⋮									

The research data were created by linking the household dwelling unit to the sampling unit and subsequently adding the household income information of the dwelling unit. This procedure was iterated for all waves of the panel so that the changing household composition was taken into account. All analyses use register information for respondents and nonrespondents alike.

The dataset of 2 500 sampling units includes also persons not belonging to the study population and hence the total sample size was slightly reduced. However, even after the exclusion of the overcoverage there were still persons creating difficulties in the analyses: for some households there were no household income information available in the household registers of Statistics Finland. For attaining the comparability between the former analysis performed for ECHP and current analysis performed for EU-SILC, persons having missing values in the income variables were excluded from the analyses of EU-SILC dataset also (Gerks 2004).

The total amount of 2 353 households had their income available in the registers in all analysed income reference years and thus it is also the starting point of the analyses of this paper. From now on this subset of the original sample is referred as the FULL-sample. The total amount of respondents to all four waves is 1 448 and the number of persons responded at the initial wave is 1 769. Thus with the FULL-sample of 2 353 respondents the amount of initial wave nonresponse is 584 persons being round 25 percent. From now on the respondents to all four waves are referred as the OBS-sample (obs as observed) and respondents at the initial wave are referred as the RESP-sample (resp as respondents).

2.1 Quintiles of disposable household equivalised income

The main attention of this paper is given to analysis of transitions between income quintiles of the disposable household equivalised income. The transition analysis bases on the FULL-sample and hence it is enough to analyse the income quintiles of the FULL-sample only. As it was mentioned previously the income reference period is always the year preceding the survey year, that is income distributions of EU-SILC 2006 base on income data from reference year 2005 and so on. Income quintiles for the disposable household equivalised income are displayed in Table 2.

Table 2: The disposable household equivalised income, FULL-sample

Year	**N**	**The 20 th**	**The 40 th**	**The 60 th**	**The 80 the**
2005	2 353	12 720	17 147	21 302	28 119
2006	2 353	13 016	17 708	22 492	29 550
2007	2 353	13 668	18 907	23 854	31 392
2008	2 353	14 384	19 524	24 777	32 130

If the respondents were allocated to quintiles according to the percentile points of the analyse year, 20 % of the respondents would be in every quintile in every analyse year. This is of course the idea of quintile computation but it does not help in the analysis of transitions. Thus the percentile points of the income reference year 2005 are selected as fixed percentile points for all subsequent panel waves.

An adjustment to the percentile points is still required because of the inflation, for instance. The idea is to adjust the percentile points by the ratio of the median of the analyse year and the median of the base year 2005. Adjusting by the median ratios is a method Gerks (2004) applied in the analyses of European Community Household Panel in order to prevent the clustering of respondents into a one quintile. The medians and median ratios are displayed Table 3. The income bracket for quintile number one in 2008 is $1.14 \times 14384 = 16397.76$, for instance.

Table 3: The medians and median ratios for the FULL-sample

Year	**N**	**Median**	**Median**	**Median**
			2005	**ratio**
2005	2 353	19 322	19 322	1.00
2006	2 353	19 956	19 322	1.03
2007	2 353	21 326	19 322	1.10
2008	2 353	22 054	19 322	1.14

3 Markov chains

3.1 Theory

In general, let $\{X_t, t = 0, 1, 2, \ldots\}$ be a discrete time stochastic process with finite state space $E = \{0, 1, \ldots, N\}$. If the conditional probabilities at time $t + 1$ satisfy

$$P(X_{t+1} = j | X_0 = i_0, X_1 = i_1, \ldots, X_t = i) = P(X_{t+1} = j | X_t = i), \tag{1}$$

the stochastic process is called a Markov chain. The property 1 is also known as the Markov property or the memoryless property. (Brémaud 1999)

The conditional probabilities are called transition probabilities and they are collected into the matrix $\mathbf{P} = \{p_{ij}\}_{i,j,\in E}$, where

$$P(X_{t+1} = j | X_t = i) = p_{ij}. \tag{2}$$

If the probabilities defined in equation 2 are independent of the time point t, the Markov chain is said to be time homogeneous. The transition probabilities sum to one, i.e. $\sum_{j=1}^{k} p_{ij} = 1$, where $p_{ij} \geq 0$ for all $i, j \in 1, \ldots, k$. (Brémaud 1999)

The random variable X_t at time point $t = 0$ is called the initial state of the Markov chain with initial probability distribution ν given by

$$P(X_0 = i) = \nu(i), \tag{3}$$

where $i \in E$ and $\sum_{i=1}^{k} \nu(i) = 1$. The initial distribution tells us how the Markov chain starts. The initial distribution and the transition matrix determine the distribution of the discrete-time homogeneous Markov chain. (Brémaud 1999)

The state of the Markov chains after t-steps is computed by using the t-step transition probabilities defined by

$$p_{ij}^{(t)} = P(X_{s+t} = j | X_s = i), \tag{4}$$

where $s \geq 0$ is the selected time point and t is the selected time interval or time step. Because of the homogeneity assumption of the chain, probabilities in equation (4) are independent of the value of s. (Häggström 2002)

The t-step transition probabilities are computed from the first step transition probabilities and are given by

$$p_j^{(t)} = \sum_{i \in E} \nu_i p_{ij}^{(t)}, \tag{5}$$

where $p_j^{(t)} = P(X_t = j)$ is the probability of being at state j after t time steps and $\nu_i = P(X_0 = i)$ is the initial probability of state i. The initial distribution, the distribution after t time steps and the transition probabilities are possible to present in matrix form. Hence equation (5) becomes

$$\mathbf{p}^{(t)} = \boldsymbol{\nu}\mathbf{P}^t, \tag{6}$$

where $\mathbf{p}^{(t)} = (p_0^{(t)}, p_1^{(t)}, \ldots, p_N^{(t)})$ is the t-step distribution, $\boldsymbol{\nu} = (\nu_0, \nu_1, \ldots, \nu_N)$ is the initial distribution and $\mathbf{P}$ is the matrix containing first step transition probabilities. (Brémaud 1999)

3.2 Computation of transition probabilities

Markov chain modelling is used for studying transitions between adjusted income quintiles. A person (representing his/her household) belonging to the FULL-sample has five possible income quintile states in every analyse year. Thus the Markov chain in question is a process $\{X_t\}$, where $t = 0, 1, 2, 3$ with a state space consisting of the quintiles $E = \{1, 2, 3, 4, 5\}$. Here quintile one is assigned to the lowest income class and respectively quintile five is assigned to the highest income class. If it is also supposed that the time has no effect on the first step transition probabilities, the process in question is a discrete time and a finite discrete state space homogeneous Markov chain.

Because the Markov chain in question satisfies the conditions of equation (1) the transition probabilities are computed by applying equation (2). The actual calculation is done with the ℓEM-software in which the estimation bases on maximum likelihood theory and on the EM-algorithm (Vermunt 1997). The estimated transition probabilities are displayed in Table 4. The transition probabilities are supposed to be the same for respondents and nonrespondents.

As expected transitions into the current state are more probable than into any other states. This phenomenon is perhaps greater in the top and bottom quintiles, where the probability at staying in the current state is over 75 %. This high probability at staying in the current state may have something to do with the fact that a household belonging to the lowest income quintile and having a decreasing amount of income every year cannot change its quintile position. The same applies to the uppermost quintile class when the word decreasing is replaced by increasing.

Table 4: The estimated transition probabilities

	End				
Start	1	2	3	4	5
1	0.7586	0.1664	0.0454	0.0213	0.0083
	(0.0112)	(0.0098)	(0.0055)	(0.0038)	(0.0024)
2	0.1601	0.5779	0.1863	0.0575	0.0182
	(0.0099)	(0.0133)	(0.0105)	(0.0063)	(0.0036)
3	0.0450	0.1709	0.5255	0.2181	0.0405
	(0.0057)	(0.0103)	(0.0137)	(0.0113)	(0.0054)
4	0.0258	0.0580	0.1599	0.6013	0.1550
	(0.0042)	(0.0062)	(0.0097)	(0.0129)	(0.0096)
5	0.0300	0.0137	0.0348	0.1331	0.7884
	(0.0045)	(0.0030)	(0.0048)	(0.0089)	(0.0107)

4 Results

The results of the empirical and Markov chain analyses are collected into Table 5. The transition probabilities of the FULL-sample displayed in Table 4 are used for attaining the Markov chains results also for the RESP-sample. It seems that there is only small amount of initial wave nonresponse bias present in the analysed rotation group of Finnish EU-SILC. Because of lacking nonresponse at the initial wave there is no fade-away phenomenon present in the dataset. It is clear that attrition biases the results making the distribution into quintiles more skewed during the lifetime of the panel.

Table 5: Empirical and theoretical analysis results

	Year 2005		**Year 2008**				
	Full	**Resp**	**Full**		**Resp**		**Obs**
Sample size	2 353	1 769	2 353		1 769		1 448
Distr. on states	**Emp**	**Emp**	**Markov**	**Emp**	**Markov**	**Emp**	**Emp**
$p(1)$	20.0	19.3	20.8	20.4	20.5	20.5	18.9
$p(2)$	20.0	20.1	19.4	19.8	19.3	19.3	18.7
$p(3)$	20.0	20.0	18.4	18.7	18.4	18.2	18.1
$p(4)$	20.0	20.5	20.9	21.1	21.0	21.7	22.2
$p(5)$	20.0	20.1	20.6	20.1	20.7	20.4	22.1

From the Markov chain modelling point of view comparisons made between the simulated and empirical distributions in Table 5 are promising. The simulated distributions for the FULL- and RESP-samples are close to their empirical distributions and hence the usage of the Markov chain modelling is justified.

The computed transition probabilities could be used for simulating the future states of the chain to see what happens to the initial wave nonresponse bias with longer duration of the panel than four years. But because the selected sample contains no real initial wave nonresponse bias it is clear that simulation will not provide any support to the fade-away hypothesis. Hence the simulation results are not displayed here and the reader is advised to consult Junes (2012, p. 57) for the results.

References

Atkinson, T., Cantillon, B., Marlier, E. & Nolan, B. (2002). *Social Indicators: The EU and Social Inclusion*. Oxford University Press.

Brémaud, P. (1999). *Markov Chains: Gibbs Field, Monte Carlo Simulation, and Queues*. New York: Springer.

Gerks, H. (2004). *Zur Stabilität von Nonresponse-Effekten in Panelerhebungen.* Bachelor's thesis, Freie Universität Berlin.

Häggström, O. (2002). *Finite Markov Chains and Algorithmic Applications.* Cambridge University Press.

Junes, T. (2012). *Initial Wave Nonrespone and Panel Attrition in the Finnish Subsample of EU-SILC.* Master's thesis, University of Helsinki.

Rendtel, U., Behr, A., Bellgardt, E., Neukirch, T., Pyy-Martikainen, M., Sisto, J., Lehtonen, R., Harms, T., Basic, E. & Marek, I. (2004). Report on Panel Effects. Results of Work Package 6 of the CHINTEX-Project, CHINTEX. Retrieved 11.1.2012 from `http://www.destatis.de/jetspeed/portal/cms/Sites/destatis/Internet/DE/Content/Wissenschaftsforum/Chintex/ResearchResults/Downloads/WorkingPaper22,templateId=renderPrint.psml`.

Vermunt, J. K. (1997). ℓem: A General Program for the Analysis of Categorical Data. The Netherland: Tilburg University. Retrieved 3.5.2012 from `http://spitswww.uvt.nl/web/fsw/mto/lem/manual.pdf`.