Documentation of ESS Post-Stratification Weights

25th April 2014

http://www.europeansocialsurvey.org/

Contents

1 Introduction

Post-stratification weights are a weighting system that uses auxiliary information to reduce the sampling error and potential non-response bias, in comparison to pure design based weights. They have been constructed using information on variables for age, gender, education, and region. The post-stratification weights are obtained by adjusting the design weights in such a way that they will replicate the distribution of the so called control data. As control data the two marginal population distribution have been used, one for the cross-classification of *age* (i.e age classes), *gender*, and *education* (GAE) and the second one for the variable *region*. The population distributions for those adjusting variables are obtained from the European Union Labour Force Survey (LFS) data. The advantage of post-stratification weights over design weights is that:

- They can reduce the sampling error, if it can be expected that there is some (linear) dependency between the variable of interest and the variables used for post-stratification.
- They can reduce an existing non-response bias if there is a (linear) dependency between response and the variables used for post-stratification.

2 Methodology

For most countries and rounds there is control data for gender, age, education and region. However, given the available control data it was not in all cases deemed appropriate to adjust the weighted sample data to the joint distribution of gender, age, education and region. Hence, it was decided not to use a straightforward post-stratification, but a raking procedure instead. The raking procedure uses iterative post-stratification to match weighted marginal distributions of a sample to known population margins. The software used to calculate the weights is `R` (R Core Team, 2013) applying the `survey` package, (Lumley, 2013). The technique is similar to iterative proportional fitting, here post-stratification is applied iteratively for the known population margins given the post-stratification weights of the step before until a convergence is reached, i.e. the weights stop changing (see, Lumley, 2010, page 139). There are some exceptions from this procedure, Table 1 gives an overview on the used adjustment variables by country. It should be noted that for those cases where only one adjustment variables is used, e.g. only GAE, we have in fact post-stratifications weights. The post-stratification weights have then the property that,

$$\sum_{i \in s} \frac{w_i x_i}{\sum_{i \in s} w_i} = \overline{x} \, ,$$

where s is the net sample, w_i is the post-stratification weight and x_i is the observation of adjustment variable x, e.g. an indicator for a GAE class or region, of the i-th element in s. Finally, $\overline{x}$ the population mean of x.

3 The Control Data

The major source for the control data is the LSF provided by Eurostat. The decision is predominantly due to good and continuous coverage of the data, but also because national LFS teams are typically relatively large and they have the expertise with which they clarify the methodological issues around population controls with Eurostat. Exact sources of the data are listed in Table 2.

3.1 Gender

For gender, the „gndr“ variable in the ESS datasets is recoded to:

0= Missing.

1= Male.

2= Female.

3.2 Age

For age, the „agea“ variable in ESS datasets is recoded to:

0 = Missing.

1 = $15 - 34$ years old.

2 = $35 - 54$ years old.

3 = 55+ years old.

There is a small issue with age in LFS where data for age groups above 75 years is not provided, i.e. Iceland, Norway, and Sweden (see, Table 2). We addressed the problem of missing population data for older population by incorporating control data (i.e. margin for age for 75+) from the ESS appendix, which has complete data for age. However, with that we then lack full interaction with education. Thus, a modified weighting approach will be used for these three countries.

3.3 Education

For education, the „edulvla“ variable in ESS datasets in ESS data is recoded to „edulvlvR“:

0 = Missing.

1 = Lower education (lower secondary or less) includes ISCED „level 0 Not completed primary education“, „1 Primary or first stage of basic“, and „2 Lower secondary or Second stage of basic education“. Also short vocational programs (less than 3 years) taken after primary school (shorter 3C programs), labeled in LFS with „22“.

2 = Medium education (higher secondary and post-secondary, non-tertiary) includes ISCED level „3 Upper secondary (A, B, C)“ and „4 Post-secondary, non-tertiary“.

3 = Higher education (post-secondary) includes ISCED level 5 and higher levels, i.e. any stage of tertiary education (e.g. BA, BSc, MA, PhD), including vocational ISCED 5B programs which have different names in different countries.

3.4 Region

In contrast to education which is standardized to three levels, each country has a different region variable which varies in the number of categories. All control data (LFS) are given at NUTS2 level (Eurostat NUTS), while some countries in the ESS use different classifications: NUTS1 (less detailed), NUTS2, NUTS3 (more detailed) or sometimes even partially aggregated NUTS2 or NUTS3 classifications (Switzerland, Greece, Portugal and Ukraine). For weighting purposes the region data is recoded to common denominator so that ESS and LFS categories match as presented by Table 3. When control data (NUTS2 level) has more categories than ESS data, the former is usually recoded to NUTS1 level. When ESS sample data is more detailed (usually NUTS3) than LFS control data (NUTS2), then the former is recoded to NUTS2 level. In some instances to another common denominator, which is actually a partial aggregation of NUTS2 into a lower number of categories. This is needed when one of the rounds has a different number of categories than others (Switzerland, Greece, Finland), or, when some regions are excluded (Portugal, France), or a non-NUTS coding used in population data (Ukraine).

4 Strategy for Dealing with Missing Data

Control variables, especially education, can have a lot of missing values on sample and on control data. This is an issue, particularly for the GAE variable. Table 4 gives an overview on how missing values in the control and/or sample data have been handled regarding the GAE variable.

There are three situations how missing data can appear in GAE tables:

1. Missing value only in sample cell. Values in missing cells are copied to corresponding cells in control data table, taking the missing at random assumption (MAR). Next, other cells in the control data table are proportionally adjusted so that the total sum and the ratio between existing cells are preserved.

2. Missing value only in control data. Usually, we ignore them, using the so called missing completely at random assumption (MCAR). However, if unknown values present more than 1 % of the population, assuming MCAR is risky as population

structure could be affected. In these cases there is another alternative, assuming an equal distribution of unknown values, i.e. missing at random (MAR), and equally re-allocating them between known values. In Table 4 countries with MCAR assumption are labeled with I, while those with MAR assumption are labeled II.

3. Missing values both in sample and in control data. If the missing value in sample is lower than in the control, then the cell is normally used. On the other hand, cells where the missing value on control data is substantially higher than on sample data are treated in a similar way as in cases with missing value only in control data (item 2 above). The control value was decreased to the sample value and the equally re-allocated among other values assuming missing at random (MAR). The MAR assumption in Table 4 is labeled with III.

5 Weight Truncation and Scaling

As with the design weights also the post-stratification weights are scaled to the sample size, i.e. the initial weights provided by the post-stratification procedure are divided by their arithmetic mean. Then weights are truncated around the value of 4.

References

Lumley, T. (**2010**): Complex Surveys, A Guide to Analysis Using R. Hoboken: Wiley.

Lumley, T. (**2013**): survey: analysis of complex survey samples. R package version 3.29. URL http://CRAN.R-project.org/package=survey

R Core Team (**2013**): R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/

A Tables

Table 1: Used Adjustment Variables for Post-Stratification

	Country	Control Variables			#Regions
1	AT	GAE	R	-	9
2	BE	GR	AR	ER	3
3	BG	GAE	-	-	6
4	CH	GA	R	-	7
5	CY	GAE	-	-	1
6	CZ	GAE	R	-	8
7	DE	GAE	R	-	16
8	DK	GAE	R	-	5
9	EE	GAE	R	-	7
10	ES	GAE	R	-	16
11	FI	GAE	R	-	4
12	FR	GAE	R	-	9
13	GR	GAE	R	-	10
14	HR	GA	R	-	3
15	HU	GAE	R	-	7
16	IE	GA	R	-	2
17	IL	GA	E	R	7
18	IS	GA	E	-	1
19	IT	GAE	R	-	5
20	LU	GAE	-	-	1
21	LT	GAE	-	-	1
22	LV	GAE	-	-	1
23	NL	GAE	R	-	12
24	NO	GAE	R	-	7
25	PL	GA	R	-	16
26	PT	GAE	R	-	5
27	RO	GAE	R	-	8
28	RF	GAE	R	-	10
29	SE	GAE	R	-	8
30	SI	GAE	R	-	2
31	SK	GAE	R	-	4
32	TR	GAE	R	-	12
33	UA	GA	R	-	11
34	UK	GA	R	-	12

GAE	Cross-classification of the gender, age, and education variable
GA	Cross-classification of the gender and age variable
R	Region variable
E	Education variable
AR	Cross-classification of the age and region variable
ER	Cross-classification of the education and region variable

Table 2: Source of Control Data and Eventual Corrections Done in ESS Sample Data

	Country	ESS1	ESS2	ESS3	ESS4	ESS5
1	AT	LFS	LFS	LFS	-	-
2	BE	LFS	LFS	LFS	LFS	LFS
3	BG	-	-	LFS	LFS	LFS
4	CH	LFS	LFS	LFS	LFS	LFS
5	CY	-	-	LFS	LFS	LFS
6	CZ	LFS	LFS	-	LFS	LFS
7	DE	LFS	LFS	LFS	LFS	LFS
8	DK	LFS	LFS	LFS	LFS	LFS
9	EE	-	LFS	LFS	LFS	LFS
10	ES	LFS	LFS	LFS	LFS	LFS
11	FI	LFS APP	LFS APP	LFS APP	LFS APP	LFS APP
12	FR	LFS	LFS	LFS	LFS	LFS
13	GR	-	LFS	-	LFS	LFS
14	HR	-	-	-	LFS APP	LFS APP
15	HU	LFS	LFS	LFS	LFS	LFS
16	IL	APP	-	-	APP	APP
17	IS	-	LFS	-	LFS APP (75+)	- -
18	IE	LFS	LFS	LFS	LFS	LFS
19	IT	LFS	LFS	-	-	-
20	LU	LFS	LFS	-	-	-
21	LT	-	-	-	LFS (EDU ADJ)	-
22	LV	-	-	LFS	LFS	-
23	NL	LFS	LFS	LFS	LFS	LFS
24	NO	LFS APP (75+)	LFS APP (75+)	LFS APP (75+)	LFS APP (75+)	LFS APP (75+)
25	PL	LFS	LFS	LFS	LFS	LFS
26	PT	LFS	LFS	LFS	LFS	LFS
27	RO	-	-	LFS	LFS	-
28	RF	-	-	APP (EDU ADJ)	APP (EDU ADJ)	APP (EDU ADJ)
29	SE	LFS	LFS	LFS	LFS	LFS APP (75+)
30	SI	LFS	LFS	LFS	LFS	LFS
31	SK	-	LFS	LFS	LFS	LFS
32	TR	-	LFS (R4)	-	LFS	-
33	UA	-	APP	APP	APP	APP
34	UK	LFS	LFS	LFS	LFS	LFS

-	Country did not participate in that round
LFS	European Union Labour Force Survey
APP	Various sources in ESS Appendix Population Statistics
(75+)	LFS used in general, except for data misses age category (75+) (see section 2.2.1)
(EDU ADJ)	Sample data needed adjustment to match control data (see section 2.2.2)
(R4)	No data for particular round, neighboring round used instead

Table 3: Recoding of the Region Variable

	Country	ESS	RECODE	Notes, Exceptions, etc.
1	AT	NUTS2	/	Only change order of precedence.
2	BE	NUTS1	NUTS2 to NUTS1.	R5 matches NUTS2 but we still recode it to NUTS1.
3	BG	NUTS3	NUTS3 to NUTS2.	
4	CH	NUTS2	NUTS2 to 6 regions.	Only R1 has 6 regions (later rounds match NUTS2) but the recoding is done for all for better comparison.
5	CY	NUTS3	NUTS3 to NUTS2.	
6	CZ	NUTS3	NUTS3 to NUTS2.	Except for R4 when NUTS2 is already used in ESS.
7	DE	NUTS1	NUTS2 to NUTS1.	Note that in R1 there were 33 regions in NUTS2.
8	DK	NUTS3	NUTS3 to NUTS2.	Except for R4 and R5 when NUTS2 is already used in ESS. There is no LFS data for R1, R2, and R3 so we use R4 data.
9	EE	NUTS3	NUTS3 to NUTS2.	
10	ES	NUTS2*	NUTS2 to NUTS1.	We recode to NUTS1 (16 regions) because of certain empty cells in ESS region data.
11	FI	NUTS3*	NUTS2 to 4 regions.	R1 (5 regions but not NUTS2) and R5 (19 regions, NUTS3) are also recoded to 4 regions. Region variable taken from ESS Appendix instead of LFS.
12	FR	NUTS1	NUTS2 to NUTS1.	R5 matches NUTS2 but we still recode it to NUTS1.
13	GR	NUTS2*	NUTS2 to 10 regions.	Because in R4 we have only 10 regions we recode also all other rounds to 10 regions.
14	HR	NUTS2	/	Weighting possible without recoding.
15	HU	NUTS2	NUTS3 to NUTS2.	Recoding needed only for R5 (NUTS3, 20 regions). For others only change order of precedence and labels.
16	IE	NUTS3	NUTS3 to NUTS2.	Different number of regions in each round (3 in R3, 4 in R4, 8 in others). Recode all to NUTS2.
17	IL	non-NUTS	/	Weighting possible without recoding.
18	IS	NUTS2	/	No recoding, no weighting with region.
19	IT	NUTS2*	NUTS2 to NUTS1.	We recode to NUTS1 (5 regions) because of certain empty cells in ESS region data.
20	LT	NUTS3	NUTS3 to NUTS2.	
21	LU	NUTS2	/	No recoding, no weighting with region.
22	LV	NUTS3	NUTS3 to NUTS2.	
23	NL	NUTS3	NUTS3 to NUTS2.	
24	NO	NUTS2	/	Weighting possible without recoding.
25	PL	NUTS2	/	Only change order of precedence.
26	PT	NUTS2*	NUTS2 to 5 regions.	In ESS two island regions (Azores & Madeira) are excluded. Change also order of precedence.
27	RO	NUTS2	/	Weighting possible without recoding.
28	RU	non-NUTS	/	Weighting possible without recoding.
29	SE	NUTS2	NUTS3 to NUTS2.	Recoding needed only for R5 (NUTS3, 21 regions). For others only change order of precedence.
30	SI	NUTS3	NUTS3 to NUTS2.	Note that in R5 a new classification is used with 16 regions.
31	SK	NUTS3	NUTS3 to NUTS2.	
32	TR	NUTS1	NUTS2 to NUTS1.	
33	UA	non-NUTS	non-NUTS to 11 regions	We recode to NUTS1 (11 regions) because of certain empty cells in ESS region data.
34	UK	NUTS1	NUTS2 to NUTS1.	

Table 4: Handling of Missing Values in the Control and Sample Data

	Country	Structure of missing values in GAE table					Procedure for handling missing values				
		R1	R2	R3	R4	R5	R1	R2	R3	R4	R5
1	AT	S	S+(P)	S	-	-	I	I	I	-	-
2	BE	NoE	NoE	NoE	NoE	NoE	I	I	I	I	I
3	BG	-	-	S	N	S	-	-	I	I	I
4	CH	NoE	NoE	NoE	NoE	NoE	I	I	I	I	I
5	CY	-	-	S	S	S	-	-	I	I	I
6	CZ	S+(P)	S	-	(P)	(P)	I	I	-	I	I
7	DE	S+P *	S+P *	S	S+(P)	S+(P)	III	III	I	I	I
8	DK	S+P	S+(P)	N	(P) *	P *	II	I	I	III	III
9	EE	-	P	S+P	S+P *	S+P *	-	III	II	III	II
10	ES	S+(P)	S	S+(P)	S+(P)	S+P	I	I	I	I	II
11	FI	S	S	S	S	S	I	I	I	I	I
12	FR	S+(P)	S	N	S	S+(P)	I	I	I	I	I
13	GR	S	S	-	S	S	I	I	-	I	I
14	HR	-	-	-	NoE	NoE	-	-	-	I	I
15	HU	S *	S+P *	S+P	S+P	S+P	III	II	III	II	II
16	IE	NoE	NoE	NoE	NoE	NoE	I	I	I	I	I
17	IL	NoE	-	-	NoE	NoE	I	-	-	I	I
18	IS	-	NoE	-	-	-	-	I	-	-	-
19	IT	S	S	-	-	-	I	I	-	-	-
20	LU	S	S+(P)	-	-	-	I	I	-	-	-
21	LT	-	-	-	N	-	-	-	-	I	-
22	LV	-	-	S+(P) *	P	-	-	-	III	II	-
23	NL	S+(P)	S+(P)	(P)	(P)	(P)	I	I	I	I	I
24	NO	N	S+(P)	S	P *	P *	III	I	I	III	III
25	PL	NoE	NoE	NoE	NoE	NoE	I	I	I	I	I
26	PT	N	S	S	S	S	I	I	I	I	I
27	RO	-	-	S	S	-	-	-	I	I	I
28	RF	-	-	S+P	S+P	S+P	-	-	I	I	I
29	SE	S+P *	P *	S+P *	P *	(P) *	III	III	III	III	I
30	SI	S *	S+(P)	S	S	S	III	I	I	I	I
31	SK	-	S	S	S	S	-	I	I	I	I
32	TR	-	S	-	S+(P)	-	-	I	-	I	-
33	UA	-	NoE	NoE	NoE	NoE	-	I	I	I	I
34	UK	NoE	NoE	NoE	NoE	NoE	I	I	I	I	I

NoE	no 3-dimensional GAE table (education is in a separated table)
N	no issues (correspondence between missing value cells on sample and control data)
S	there are missing value cells on sample that do not exist in control data
P	there are missing value cells on control data that do not exist in sample data
()	missing values do not exceed 1%
*	missing value on control data for at least 1 point higher than on sample
I (MCAR)	Standard procedure for handling missing values used in most countries
II (MAR)	Re-allocation procedure for handling missings for countries with more than 1% of missings in LFS data
III (MAR)	Re-allocation procedure for handling missings for countries with more than 1% of missings in LFS data and large discrepancy from the ESS missing structure