Short-Documentation of the Update of the SOEP-Weights, 1984–2008

Martin Kroh
DIW Berlin, Socio-Economic Panel Study
Mohrenstr. 58, D–10117 Berlin, mkroh@diw.de

1 Introduction

With the 2008 data, we are distributing a revised version of the SOEP weighting variables from waves 1984 to 2008. The amendments affect the cross-sectional weights (`AHHRF-YHHRF`) only and not the longitudinal weighting variables (`BHLEIB-YHBLEIB`). Several developments led to the retroactive update of the cross-sectional weights.

First, the SOEP data can now be linked to extensive regional information that was unavailable at the time of the initial sampling. A growing number of commercially provided neighborhood characteristics can be linked to addresses of SOEP households (MICROM). Moreover, the census office is providing a growing number of regional characteristics on the county level in Germany (INKAR - Indikatoren und Karten zur Raum- und Stadtentwicklung). We use these retrospectively collected regional data to improve the fit of first-wave response models, which –as with any cross-sectional survey– suffer from the paucity of information available on both respondents and non-respondents.

Second, in the past, the strategy of post-stratifying the weighting variables to correspond with certain marginal distributions of the overall population (age, gender, household size, etc.) as recorded by the German census was introduced gradually over time. In the update of the cross-sectional weights, we use the same set of population characteristics more consistently for post-stratification in all waves of the SOEP and also extend post-stratification to match the regional distribution of households in Germany (size of communities and states (=Bundesländer)).

Finally, our impression from users' feedback was that after 25 waves, the growing number of weighting variables for each wave but also for different combinations of sub-samples (e.g., one set omitting certain sub-samples, one

set ignoring first-wave respondents, etc.) made the SOEP less accessible to new users. One aim of the revision was thus to concentrate on the "standard" variables in the data distribution.

The present report is preliminary in that it provides information only about the most important changes in how the SOEP cross-sectional weights are derived, but does not, for instance, provide each estimate of the underlying response models. We are currently compiling more comprehensive documentation of the SOEP weighting variables that will contain information on the sampling design, the attrition analysis in each wave, and the derivation of the cross-sectional weights. The derivation of the longitudinal weighting variables in wave Y (2008) is documented, just like in the previous waves, in the DIW Research Note 47 entitled "Documentation of Sample Sizes and Panel Attrition in the German Socio Economic Panel (SOEP) (1984 to 2008)".

The first part of this brief report describes changes in the first-wave response models for Samples A, B, E, F, G, and H. The section thereafter lists the characteristics of the population living in Germany reported by the national census bureau that were used to post-stratify the SOEP cross-sectional weighting variables in all waves. Finally, the report contains the new list of available weighting variables.

2 Revision of First-Wave Weights

In contrast to models of non-response in waves 2 and following, the informational basis to estimate such a wave 1 is often limited. In waves 2 and following, one can draw on the wealth of information collected in the previous wave(s) to correct for non-response bias, while in wave 1 one often lacks information on units of analysis that elect not to participate. Table 1 gives an overview of sources of information available for each of the SOEP's gross samples A though H in their respective first-waves.

The stratification of samples provides a set of basic information on both respondents and non-respondents. For instance, sample B is stratified by migrants' countries of origin (Turkish, former Yugoslavian, Italian, Greek, and Spanish). On this basis, one can regress non-response on nationality. Sample G was selected from a much larger screening sample of households that responded to a first interview to determine whether or not they belonged to the target population of high-income households. Basic demographic information from the screening phase can again be used to estimate a response model. With the exception of the most recent sample H, all sub-samples contain proxy information reported by the interviewer. For instance, interviewers

reported on characteristics of the sampled address, the residential environment, and the larger neighborhood. Also, interviewers were instructed to collect some basic information on the household (household size, gender, and age of person answering the door, etc.).

Table 1: Gross-Sample Information Available for the Different Sub-Samples

Sample	A	B	E	F	G	H
	1984		1998	2000	2002	2006
Household Information						
Sample Stratification		+	+	+	+	
Infor. on Household by Screening Survey					+	
Residential Infor. by Interviewer	+	+	+	+	+	
Infor. on Household Members by Interv.	+	+	+	+	+	
Regional Information						
Location (state, district size)	+	+	+	+	+	+
County Characteristics (INKAR)			+	+		+
Neighborhood Characteristics (MICROM)						+

Regional information on the size of the community and the state (Bundesland) in which the sampled address is located are available in all sub-samples, but only the more recent samples make it possible to consider externally provided regional information on a smaller scale in the non-response analysis. For samples E, F, and H, county-level information can be merged to the respective gross-samples, which cover a wide range of regional characteristics, such as demographic development, prices, wealth, occupational structure, and industries, etc.[1] At the lowest regional level –that of the household

[1]In particular, non-response models consider unemployment rate, the movement of the unemployment rate, percentage of single-family homes, living area, labor force participation rate, trends in the labor force participation rate, female labor force participation rate, percentage employed in the primary sector, percentage employed in the secondary sector, percentage employed in the tertiary sector, percentage of the population below 5 years of age, percentage of the population between 6 and 17, percentage of the population between 18 and 24, percentage of the population between 25 and 29, percentage of the population between 30 and 49, percentage of the population between 50 and 64, percentage of the population over 65 years of age, household size, migration balance, natural balance, male life expectancy, female life expectancy, employee pay, household income, residents per doctor, tax revenues, trends in tax revenues, population density, rate of welfare participation, per capita GDP, trends in per capita GDP, building land prices, and trends in building land prices.

address– the first-wave response model for sample H draws on recently released neighborhood data that, again, cover a wide range of socio-economic and demographic characteristics.[2] It was mainly the recent release of external information on county and neighborhood characteristics covering the past 15 years that led to our revision of first-wave weights in SOEP.

3 Revision of the Post-Stratification

The revised cross-sectional weights of the SOEP are uniformly adjusted to the same set of characteristics of the underlying population in all years. At the household level (`AHHRF-YHHRF`), we consider census information on the number of households per state (i.e., Bundesland),[3] district magnitude,[4] household size,[5] and home ownership status. Individual-level weights (`APHRF-YPHRF`) match the margins of the census bureau with respect to age groups,[6] gender, and the number of inhabitants with non-German nationality.

4 Weighting Variables in the SOEP

The wide-format data files `HHRF` for the household level and `PHRF` for the individual level contain the longitudinal 'inverse staying probabilities' (`BHBLEIB-YHBLEIB`) for each wave. Moreover, in the previous data distributions, for each cross-section we included one set of weighting factors that excluded sample G, one set of weighting factors that excluded first-wave respondents, and one set of weighting factors that included both sample G and first-wave respondents. Moreover, for each wave we delivered specific cross-sectional weights for samples D and G.

[2]The social-structural composition of the neighborhood in terms of wealth, families, age, migration background, anonymity, and life style typologies; the mobility in the neighborhood in terms of the migration volume, migration balance, short-distance migration, long-distance migration, and also the housing and street typology. Moreover, MICROM delivers information on the financial characteristics of a neighborhood, such as credit card use, installment payments, building society savings, and money investments.

[3]To avoid cells with too few observations in post-stratification, we do not consider city-states (Berlin, Hamburg, and Bremen) and Saarland separately but instead merge them with their neighboring states (Berlin/Brandenburg, Lower Saxony/Bremen, Rhineland-Palatinate/Saarland, and Schleswig-Holstein/Hamburg).

[4]We distinguish between communities below 20,000 inhabitants, between 20,000 and 100,000, between 100,000 and 500,000, and over 500,000 inhabitants.

[5]That is, the number of single-person households, two-, three-, four-, and five- or more person households.

[6]Distinguishing between age groups 0-15, 15-20, 20-25, 25-30, 30-35, 35-40, 40-45, 45-50, 50-55, 55-60, 60-65, and 65 or older.

To facilitate access to SOEP weights for new users, we have reduced the number of variables in comparison to previous data distributions. In the update of the data files, we only consider one set of cross-sectional weighting variables labeled `AHHRF-YHHRF` at the household level and `APHRF-YPHRF` at the individual level that cover all samples A though H.

For those users who want to investigate the effect of integrating refreshment samples on their analyses, for each year in which new samples were drawn from populations that were already covered by 'old' samples we also provide additional weights for the new and the old sample(s). For instance, sample H drawn in 2006 is a representative sample of the population living in Germany as are the existing samples A through G. The 'standard' cross-sectional weight in 2006, `WHHRF`, covers all samples A though H. `WHHRFAG`, but only considers the old samples A through G in 2006 and `WHHRFH` only covers the new sample H.[7] This should enable users to compare their estimates in one year between old and new samples and also to ignore a particular refreshment sample in their weighted analysis if they wish.

```
BHBLEIB     'Inverse Staying Probability Wave 1985'/
CHBLEIB     'Inverse Staying Probability Wave 1986'/
DHBLEIB     'Inverse Staying Probability Wave 1987'/
[...]
WHBLEIB     'Inverse Staying Probability Wave 2006'/
XHBLEIB     'Inverse Staying Probability Wave 2007'/
YHBLEIB     'Inverse Staying Probability Wave 2008'/
AHHRF       'Weighting Factor Wave 1984'/
BHHRF       'Weighting Factor Wave 1985'/
CHHRF       'Weighting Factor Wave 1986'/
[...]
WHHRF       'Weighting Factor Wave 2006'/
XHHRF       'Weighting Factor Wave 2007'/
YHHRF       'Weighting Factor Wave 2008'/
OHHRFAD     'Weighting Factor Samples A-D Wave 1998'/
OHHRFE      'Weighting Factor Sample E Wave 1998'/
QHHRFAE     'Weighting Factor Samples A-E Wave 2000'/
QHHRFF      'Weighting Factor Sample F Wave 2000'/
SHHRFAF     'Weighting Factor Samples A-F Wave 2002'/
SHHRFG      'Weighting Factor Sample G Wave 2002'/
WHHRFAG     'Weighting Factor Samples A-G Wave 2006'/
WHHRFH      'Weighting Factor Sample H Wave 2006'/
```

[7]The cross-sectional weight of the new sample, in this case H, contains information on the selection probabilities in the sampling design of H, the non-response analysis in wave one (see Section 2), and post-stratification (see Section 3).