\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\* A Practical Guide to Using Panel Data

\* Simonetta Longhi and Alita Nandi

\* ISER, University of Essex

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

**Appendix A**

**The German Socio-Economic Panel (SOEP)**

This is a short introduction to the German Socio-Economic Panel (SOEP), which summarises the main characteristics of the study, also discussed in Longhi and Nandi (2014) ‘A Practical Guide to Using Panel Data’. For further details on the SOEP see the user guide referred to as the Desktop Companion (DTC):

Haisken-DeNew, John P. and Frick, Joachim R. (ed.) (2005) The Desktop Companion to the German Socio-Economic Panel (SOEP)

You can find the DTC, an excellent interactive system for extracting variables (known as SOEPinfo), FAQ, codebooks and other supporting material online at <http://www.diw.de/en/diw_02.c.221178.en/about_soep.html>

*The Survey*

The SOEP is a German household panel survey that started in 1984 and is still continuing. It is managed by the German Institute for Economic Research, DIW. The SOEP started with a sample of 16,252 individuals in 5,921 households which included an over-sample of 1,393 ethnic minority households, with heads of household who were at that time Turkish, Greek, Yugoslavian, Spanish or Italian citizens. Other samples have been progressively added over the years to either include in the sample East Germany after unification, to over-sample specific sub-populations or for refreshment. Details of these sub-samples are in Table 1 at the end of this document. Most of the sub-samples have a clustered and stratified design with over-sampling of specific sub-populations.

The SOEP is an indefinite life panel survey where interviews are conducted every year over a 12 month fieldwork period. For each sub-sample, all household members aged 16 years and over in households participating in the first wave of the sub-sample (and their children irrespective of when they were born) are eligible for interview every year as long as they live in Germany. Anyone joining these households are also interviewed and followed even if they move out of these original SOEP households.

While the sampling frame excluded institutions, if sample members move into institutions they are eligible for interview. Sample members are excluded from the sample if the entire household could not be interviewed for two successive interview years or if they firmly refused to participate further.

All interviews are conducted face-to-face. The bulk of the information comes from individual interviews of adults aged 16 and over in the interviewed households. SOEP considers the head of household to be the most well-informed person regarding the decisions and choices made by the household. Every year this same person is asked additional questions about the household such as housing condition, household expenditures and also some questions on children younger than 16 year old. Interviewers also provide additional information about the fieldwork process (for example, the number of times the household was contacted, reasons for refusal and so on) as well as some information about the household and the residential area.

*Data*

SOEP provides data in several formats including Stata, SAS and SPSS. Information collected in each wave is made available in a separate set of data files: each file corresponding to one data source. The naming convention of the data files is as follows. The name of each wave-specific file starts with wave prefixes: a, b,…, z for interview years from 1984 until 2009; from 2010 onwards these become ba, bb, bc, and so on (note that the SOEP documentation uses $ as a placeholder for wave prefix). The root name of all household level files start with the letter “h” and that for all individual level files start with the letter “p”. Substantive information collected during the interviews is available in $h, $p and $kind files (containing information about households, individual respondents and children), derived variables are in the $hgen and $pgen files and basic information about all individuals in SOEP households including children and non-respondents is in the files $pbrutto. Fixed information about every household ever interviewed and every individual in those households (including those never interviewed) such as gender, date of birth, country of birth are stored in the individual level file ppfad and in the household level file hpfad. These files also include some time-varying information such as survey status, region of residence at each interview year. Additional files provide information on retrospective employment, fertility and marital histories, biographical information on parents, children, reasons for non-participation in interview and so on. See Section 1.7, Chapters 2 and 3 of the DTC for further details about the data structure and data files.

The variable names reveal information about the year they were collected in, the unit of analysis, the questionnaire or survey instrument they were included in, the question number in that questionnaire and so on (see Table 1.5 in the DTC for the complete key). The variable names do not include any information about the substantive information they represent. As the question number of a particular question, say marital status, in the questionnaire may vary across years, the variable names do not even have the same root name across years. For example, the variable which includes marital status information across the years is ap58 in 1984, ip101 in 1992 and bcp129 in 2012. One quick way to find out the variable names is to use the interactive online search facility at <http://panel.gsoep.de/soepinfo2012/>. This online search facility can also be used to extract variables for analysis, see section 1.9 of the DTC. An alternative method of finding out variable names and extracting them is discussed in Chapter 3 of Longhi and Nandi (2014) ‘A Practical Guide to Using Panel Data’.

Sample membership is identified by the variable psample, and the interview outcome by the variable $netto. Each individual is uniquely identified within and across waves by the variable persnr and each household within a wave is identified by the variable $hhnr.

Unlike the British Household Panel Survey (BHPS) and the UK Household Longitudinal Survey (UKHLS) discussed in other parts of this online appendix, SOEP acknowledges longitudinal households: households can be followed over time, and when someone moves out of the original household, the new household is a given new household identifier but the original household identifier continues to be linked to these new households. Split off from the original household can be identified by the original household identifier, hhnr, which is provided in addition to the identifier of the current household ($hhnr). The information in $hhnr is also stored in hhnrakt which is the household counterpart of persnr. These identifier variables can be used to merge individual with household level files as well as individual or household level files across waves. For further details see Chapters 4 and 5 of Longhi and Nandi (2014) ‘A Practical Guide to Using Panel Data’.

Missing values in the SOEP are coded as -1, -2 and -3 and represent item non-response (don’t know or no answer), valid skips or not applicable, and inconsistent or implausible values, respectively. For further details see Chapters 1 and 2, Haisken-DeNew and Frick (2005).

*Identifying other household members*

The SOEP also provides indicator variables for identifying the spouse and partner, parent and twins. Other family members can be identified using the file $pbrutto, see Chapter 6 of Longhi and Nandi (2014) ‘A Practical Guide to Using Panel Data’.

*Sample Weights*

As the selection probabilities are different across the different SOEP samples and like all sample surveys, there is non-response and attrition, population estimates based on sample estimates may be biased. To correct for this, appropriate cross-sectional and longitudinal weights are provided.

The sample design for most of the samples is clustered and stratified. Clustering and stratification variables which can be used to produce correct estimates of standard errors are provided with the data.

Cross-sectional and longitudinal weights for SOEP data are available in two separate wide format files: individual weights are available in a file named phrf (which also includes the individual identifier persnr) and household weights are in file named hhrf (which also includes the original household identifier, hhnr and the current household identifier, hhnrakt). Each row of phrf is uniquely identified by persnr and each row of hhrf by hhnrakt.

Sampling and interviewer information are provided in the files called varianz.

The cross-sectional individual level weights are called $phrf and the cross-sectional household level weights are called $hhrf. Those who did not participate in a particular wave will have a zero weight for that wave. The variable $pbleib, represents the inverse of the probability that a respondent at wave $ had also responded in the previous wave. Users can then compute the longitudinal weights between any two waves by multiplying the cross-sectional weight of the earlier wave with these wave-on-wave non-response weights for the intervening waves. For further discussion on weights see Chapter 7 of Longhi and Nandi (2014) ‘A Practical Guide to Using Panel Data’, and for computation and the types of weights available in SOEP see Section 1.8, 5.3 and 5.4 in Haisken-DeNew and Frick (2005).

*History Files*

Most of the data are collected by prospective methods. Exceptions are the marital and employment initial histories which are collected when a respondent is interviewed for the first time. These are collected by retrospective method. Until 1996, some of the sample components were sometimes asked additional questions. For example, samples B and D were asked questions intended to measure integration.

SOEP collects information on the employment activity, income receipt, marital and fertility history between two interviews as well as before the respondent joined the sample. These are available in different data files described in Chapter 3 of Haisken-DeNew and Frick (2005). All these files have a similar structure: they are multi-level files where each row represents a specific spell for each respondent. For example, the file biomarsy, refers to marriage histories; each row is uniquely identified by the individual identifier, persnr and the spell number spellnr. These files generally contain information on the type of spell, the start and end dates of each spell, and whether the spell is ongoing. For further details see Chapter 12 of Longhi and Nandi (2014) ‘A Practical Guide to Using Panel Data’.

|  |  |  |  |
| --- | --- | --- | --- |
| **Table 1: Description of SOEP samples** | | | |
| **Sample Components** | **Target Population or Sampling Frame** | **Sampling Design** | **Initial Sample Size** |
| Sample A: “West German Residents” | Residents in West Germany in 1984 living in private households where the head of household was not from Turkish, Greek, Yugoslavian, Spanish or Italian ethnic groups | Multi-stage stratified and clustered sampling design; sampling frames was ADM (Arbeitsgemeinschaft Deutscher Marktforschungsinstitute) | 4,528 households |
| Sample B: “Foreigners in the FRG” | Residents in West Germany in 1984 living in private households where the head of household was of Turkish, Greek, Yugoslavian, Spanish or Italian citizenship | Over-sampled; clustered sample design; sampling frame was the immigrant registration records | 1,393 households |
| Sample C: “German Residents in the GDR” | Residents in Germany in 1990 living in private households where the head of household was a German Democratic Republic (GDR) citizen | Clustered and stratified sample design; sampling frame was the Central Residents’ File of the GDR | 2,179 households |
| Sample D: “Immigrants” | Residents in Germany in 1994/95 living in private households where at least one household member immigrated to West Germany since 1984 | For details see the DTC | 236 households in 1994 and 295 households in 1995 resulting in a total of 522 households in 1995 |
| Sample E: “Refreshment” | Residents in Germany in 1998 living in private households | Similar sampling design as Sample A (selected independently of Samples A-D) | 1,067 households |
| Sample F: “Innovation” | Residents in Germany in 2000 living in private households | Similar sampling design as Sample A and Sample E; Households with at least one household member without German nationality were over-sampled (selected independently of Samples A-E) | 6,052 households |
| Sample G: “High Income Households” | Residents in Germany in 2002 living in private households with a monthly household net income in 2002 that was 3,835 Euros or higher | Selected from Infratest-Telefon-Master Sample which was selected using a multi-stage stratified telephone sample of households (selected independently of Samples A-F) | 1,224 households |
| Note: The guidance provided in the Desktop Companion is that samples A – F should always be used together, appropriately weighted. Since 2002, for some specific variables weights are available for using samples A – G. | | | |