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

\* A Practical Guide to Using Panel Data

\* Simonetta Longhi and Alita Nandi

\* ISER, University of Essex

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

**Appendix A**

**Panel Study of Income Dynamics (PSID)**

This is a short introduction to the US Panel Study of Income Dynamics (PSID), which summarises the main characteristics of the study, also discussed in Longhi and Nandi (2014) ‘A Practical Guide to Using Panel Data’. More details can be found in:

Hill, M. (1992) *The Panel Study of Income Dynamics: A User’s Guide*, Sage.

McGonagle, K., Schoeni, R.F., Sastry, N. and Freedman, V.A. (2012) The Panel Study of Income Dynamics: Overview, Recent Innovations, and Potential for Life Coures Research. *Longitudinal and Life Course Studies* 3(2): 268-284.

And online at:

<http://psidonline.isr.umich.edu/> is the PSID webpage;

<http://www.uncg.edu/bae/people/ribar/teaching/ECO725/notes/intro_PSID.pdf> provides a short overview of PSID outlining the main features;

<http://psidonline.isr.umich.edu/Guide/FileStructure.pdf> provides details on the file structure of PSID;

<http://psidonline.isr.umich.edu/videos.aspx> provides short video podcasts explaining some key aspects of the PSID. Although these videos have been created to train interviewers, they are useful for new users as well.

The PSID is one of the oldest ongoing household panel surveys.[[1]](#footnote-1) The survey started in 1968 with a sample of about 18,000 individuals in 4,802 families living in the US, with data collected annually from the head of the family. This is the male in opposite sex married couples, and either the male or the female adult in other types of families (see McGonagle et al. 2012 for a recent overview). Hence, in contrast to other surveys,[[2]](#footnote-2) the PSID started as an individual panel survey (see also Chapter 2 of Longhi and Nandi (2014) ‘A Practical Guide to Using Panel Data’).

The respondent provides information about himself or herself and other members of the family, although in specific supplements (like the ones about retrospective histories) both members of the couple are interviewed. From the 1970s, the PSID collects details also from the spouse of the head of the family. The survey intends to follow members of the original family unit and their offspring.

The PSID includes various subsamples. The 1968 PSID sample comprises two independent sub-samples: the ‘SRC’ (Survey Research Center), and the ‘SEO’ (Survey of Economic Opportunity), referred to as the “core samples”. Originally, the SRC sample was collected by the Survey Research Center (SRC, at the Institute for Social Research, University of Michigan), while the SEO sample was collected by the Bureau of the Census for the Office of Economic Opportunity.

The SRC sample is a sample representative of the US population, with equal selection probabilities in the 48 coterminous states. In 1968 this sample produced 2,930 interviews. The SEO sample includes only people living in Standard Metropolitan Statistical Areas (SMSAs), and people living in non SMSAs in the Southern region; in this sample, low-income families are overrepresented. In 1968 this sample produced 1,872 interviews. The 18,000 individuals living in these interviewed households are permanent sample members and are said to have the “PSID gene”. These individuals and all their biological or adopted descendants are followed over time.

In 1990, an additional sample of 2,043 Latino individuals (Mexican, Cuban and Puerto Rican) was added but dropped after 1995 due to lack of funding and underrepresentation of immigrants of Asian descent (this short-lived sub-sample is rarely used by researchers for analysis). Two additional “immigrant samples” were added in 1997 and 1999.

To reduce costs, 57% of the SEO sample was also dropped in 1997; this was also the year when the interview frequency changed from annual to biennial. Interviews in the PSID were conducted face-to-face from 1968 to 1972 and by phone from 1973. This change led to much shorter interviews (20-30 minutes compared to about one hour).

Most of the information collected during the interviews relates to the household members’ current situation, with some retrospective elements.

*Opening PSID Data*

In contrast to other datasets, PSID data are not provided in Stata format. Instead, a Stata do file is provided with the dataset, which opens the data and assigns names and labels to all variables. The name of this file varies depending on the release of the data; in the 2009 release the do file was called IND2009ER.do. Hence, before being able to use the PSID data we need to open the do file provided, add the path where the PSID data are stored, save and run the do file. For more details see Chapter 3 in Longhi and Nandi (2014) ‘A Practical Guide to Using Panel Data’.

The individual-level data are structured as a wide format flat file: data for all waves are provided in the same data file which, therefore, is very large. For this reason the dataset contains an extremely large file. Not all versions of Stata might be able to open such a large number of variables. In this case the best thing to do is to modify the do file provided with the dataset. To upload in memory only the relevant variables we delete from the command infix the name (and corresponding position numbers) of the variables that we do not need for our analysis. We also need to delete the corresponding label variable commands. We can then run this modified version of the do file, which will only upload in memory the specified subset of variables.

The variable names in the PSID are not related to their content. Variables are identified by the letters ER followed by a sequence of five numbers. The variable ER30000 is the release number; variables from ER30001 to ER30866 refer to waves from 1968 to 1993 and are arranged by wave; ER31990-ER31999 and ER32001-ER32050 are summary variables such as sex, order of birth or year of birth of parents; variables from ER33101 to ER33150 refer to the 1994 Public Release Individual data; variables from ER33201 to ER33299B refer to the 1995 Public Release Individual data; and so on. The variables from ER34001 to ER34046 refer to the 2009 Public Release Individual data. It is relatively simple to identify the year to which the data refer from the label of each variable: for example, the label of the variable ER30004 is “Age of individual 68”, while the label of the variable ER30023 is “Age of individual 69”, which are the age of the individual in 1968 and in 1969. Similarly, the label of the variable ER34004 is “Age of individual 09”, which is the age of the individual in 2009. To be able to use the Stata set of commands designed for panel data, we need to “reshape” the data from wide into long format (see Section 4.4 in Longhi and Nandi (2014) ‘A Practical Guide to Using Panel Data’).

The PSID data do not include value labels: it is necessary to consult the codebook to check the meaning of the numerical values and the way missing values are coded. Generally, code 8 (or 98 or 998, etc.) represents “don’t know”, while code 9 (or 99 or 999, etc.) represents other missing data or refusal. Inappropriate/inapplicable questions (such as age of the individual in a variable referring to waves before the individual was born or entered the sample) are coded as zero.

*Combining Family and Individual-level Data*

Although the individual level data are supplied as one single wide-format file, the family files are supplied separately by year and need to be combined with the individual-level data. The best way to combine the individual and family level files is to create a long-format file for the family data, which can be merged with the long format individual-level file. To do this, first, we convert the family files into Stata format, select and rename the variables of interest and create the wave or the year identifier. We then need to rename the family identifier variable, referred to as the Family interview number, in a way that is consistent across waves and with the individual file. After that, we can merge the individual and family level files by using both the family and wave or year identifiers. For more details see Chapter 5 in Longhi and Nandi (2014) ‘A Practical Guide to Using Panel Data’.

*Identifying Family Members*

Family members in the PSID can be identified by the variables in the individual respondent file that identifies the relationship of the respondent with the head of the family. Also in this case there is one variable for each interview year. It is also worth noting that the coding of these variables changes in 1983.

Because of the length of the study, the PSID includes multiple generations. The PSID follows people when families split, or when children leave the family, thus leading to additional individual and family interviews. The PSID provides an additional file – called PID (Parents Identification) – that allows the identification of the parents, if they have ever been part of the survey. This file has been constructed using information collected from various sources about parents and children since 1993. The variables in this file contain the personal identification number of the natural and adoptive parents together with basic information about the source of the parental information. Additional information about the parents can be retrieved from the individual and family files by using their identification number.

The personal identification number of the individual in the individual respondent file can be computed by combining the variables ER3001 (1968 Interview Number) and ER3002 (Person Number); while in the parents identification file, PID, it can be computed by combining the variables PID1 (1968 Interview Number of Individual) and PID2 (Person Number of Individual). The individual respondent and the parent identification files can then be combined using this newly computed personal identification number of the respondent. For more details see Chapter 6 in Longhi and Nandi (2014) ‘A Practical Guide to Using Panel Data’.

*History Files*

The PSID provides retrospective data files for marriage, childbirth and adoptions. These files contain details of marriages, childbirth and adoptions for PSID respondents collected in the waves from 1985 onwards together with retrospective histories prior to 1985. Hence, while data for events before 1985 are collected retrospectively, the data for events from 1985 onwards are effectively collected prospectively.

The marriage file has one record per marriage. Each record includes the person identifier together with the identifier of the spouse, an identifier for the marriage, and data about timing (start and end) and circumstances of the marriage. Additional records identify those who have never been married. Linkage with the respondent file can be done using the person identifier or the identifier of the spouse. The childbirth and adoption history file has a structure similar to the marriage file and includes details of the parents, their marital status when the child was born, birthday, birth order and sex of the child, their race, and so on.

For more details see Chapter 12 in Longhi and Nandi (2014) ‘A Practical Guide to Using Panel Data’.

*Sample Weights*

As we already mentioned, not all samples of the PSID were selected with the same selection probabilities. Like all surveys, the PSID also experiences non-response and attrition of its sample members. Hence, PSID data include weights which account for differential selection probability, non-response and attrition. The data provide longitudinal individual and family weights, and from 1997 onwards, cross-sectional individual weights.

All members of the 1968 sample of households and their descendants are defined as a PSID sample family member (or as having the “PSID gene”) and always have a non-zero cross-sectional and longitudinal weight when they respond to the survey. All others who join the households of these PSID sample family members are defined as PSID non-sample family members and receive zero longitudinal but positive cross-sectional weights. Cross-sectional weights are based on longitudinal weights, where both sample and non-sample members are assigned family level weights using the “fair shares” method. Cross-sectional weights are post stratified on population characteristics using Current Population Survey data but longitudinal weights are not. It should be noted that the longitudinal weights were adjusted for cumulative panel attrition in some years only: 1969, 1974, 1979, 1984, 1989, 1993, 1997, 2003, 2007. In the intervening years, the longitudinal weights are carried over. Longitudinal family weights are the average of the individual longitudinal weights of its family members. Split-off families receive the same longitudinal weight as the original family.

Additional weights are available for specific questionnaires such as time diaries, child development supplement, and so on. A quick method of identifying PSID weight variables consists in searching for “weight” in the web-based variable search facility at <http://simba.isr.umich.edu/VS/s.aspx>.

For more details see Chapter 7 in Longhi and Nandi (2014) ‘A Practical Guide to Using Panel Data’.

For an overview of the cross-sectional weights see Heeringa, S.G., Berglund, P.A., Khan, A., Lee, S. and Gouskova, E. (2011) PSID Cross-Sectional Individual Weights, 1997-2009. Ann Arbor, Michigan, Survey Research Center, Institute for Social Research, University of Michigan.

For longitudinal weights see Gouskova, E., Heeringa, S.G., McGonagle, K. and Schoeni, R.F. (2008) Panel Study of Income Dynamics Revised Longitudinal Weights 1993-2005. Ann Arbor, Michigan, Survey Research Center, Institute for Social Research, University of Michigan

and Gouskova, E., Heeringa, S.G., McGonagle, K., Schoeni, R.F. and Stafford, F. (2009) Panel Study of Income Dynamics Construction and Evaluation of the Longitudinal Sample Weight 2007. Ann Arbor, Michigan, Survey Research Center, Institute for Social Research, University of Michigan.

All yearly user guides also provide a section on weights. These and additional technical papers on sample weights are available at <http://psidonline.isr.umich.edu/Guide/documents.aspx>.

1. The PSID does not collect information about households but about families. Here we use the terms household and family interchangeably. [↑](#footnote-ref-1)
2. For example, the German Socio-economic Panel (SOEP), the British Household Panel Study (BHPS), or ‘Understanding Society: the UK Household Longitudinal Study’ (UKHLS), which are discussed in separate sections in this online appendix. [↑](#footnote-ref-2)