

Three examples of studies employing complex sampling methodologies and producing public-use datasets

1. The Panel Study of Income Dynamics, Institute for Social Research, Survey Research Center, University of Michigan, Ann Arbor

Description: The PSID began in 1968 with a nationally representative sample of over 18,000 individuals living in 5,000 families in the United States. All participants from the 1968 sample were followed in subsequent “waves” of the study over time, regardless of where they live. All families of PSID participants are also included in the study. In 1997-1999 a sample of 511 immigrant families was added to enhance representativeness.

Study Content: Employment, Wages, Income, Expenditures, Wealth, Mortgage Distress and Foreclosures, Pensions, Philanthropy, Education, Marriage and Fertility, Health Status, Health Behaviors, Health Insurance, Program Participation, Computer Use, Housing Characteristics

Public Use Datasets: PSID data are freely accessible, with codebooks, in a variety of formats. Tutorials and other user information can be found on their website (www.psidonline.org).

Weights: Using weights provided by the PSID, it has been shown that the PSID sample continues to closely resemble the national population even after 40 years of interviewing. Information is provided about proper weighting of variables in a series of technical papers, in-person seminars and online videos. The main weight types provided are the longitudinal individual, family and cross-sectional weights.

Longitudinal weights incorporate an explicit adjustment for attrition due to non-response on a four-year cycle. Family weights are updated to reflect changes in family composition, and are the average of the positive individual weights for sampled individuals in the family and the zero-value weights for the nonsampled individuals in the family.

Cross-sectional weights are provided for estimation across questionnaires using the individual data, and allow for analysis of both participants and non-participants to estimate population level characteristics or model population characteristics at a specific time point. The cross-sectional weights are applied using the “fair shares” method, which assigns a non-zero weight to all individuals, regardless of participation. At any data collection time point, t , a non-zero cross sectional weight for each person in a PSID family can be assigned using the fair shares method as follows:

$$w_{i,t} = \sum_{i=1}^{n_f} \alpha_i \cdot w_{i,t}^*$$

where:

n_f = the total number of sample and nonsample persons in family f;

$w_{i,t}^*$ = the current non-zero individual weight for sample person, i
 = 0 if person i is nonsample;

α_i = (general) an arbitrary influence weight $\in (0,1)$, $\sum_{i=1}^{n_f} \alpha_i = 1$.

This weight then undergoes a two-step adjustment at the family level (based on the sample, age of household head, race of household head, and region of residence), and then by population totals for major demographic characteristics using the Current Population Survey (CPS) Annual Demographic Survey.

2. The National Longitudinal Surveys, U.S. Dept. of Labor, Bureau of Labor Statistics

Description: The **National Longitudinal Surveys (NLS)** are a set of surveys designed to gather information at multiple points in time on the labor market activities and other significant life events of several groups of men and women. For more than 4 decades, NLS data have served as an important tool for economists, sociologists, and other researchers. The NLSY97 cohort comprises two independent probability samples: a cross-sectional sample, and an oversample of black and/or Hispanic or Latino respondents. The cohort was selected using these two samples to meet the survey design requirement of providing sufficient numbers of black and Hispanic or Latino respondents for statistical analysis.

Study Content: Inflation and Prices, Unemployment, Productivity, Pay and Benefits, Employment, Spending and Time Use, Workplace Injuries

Public Use Datasets: NLS public use datasets and documentation are available on their website (<https://www.nlsinfo.org/investigator/pages/login.jsp>).

Weights: Weighting techniques and proper file use are discussed on their website FAQ and this link: <http://www.nlsinfo.org/nlsy97/nlsdocs/nlsy97/use97data/weights.html>. There are more than 30 pre-created weight variables in the NLSY97 dataset, with guidance for use provided for researchers on their web site. The assignment of individual respondent weights involved a number of different adjustments. Complete details are found in the *NLSY97 Technical Sampling Report*, which has step-by-step descriptions of the entire adjustment process. Some of the major adjustments are: 1.) Computation of a base weight, reflecting the case's selection probability for the screening sample. This step also corrects for missed housing units and caps the base weights in the supplemental sample to prevent extremely high weights; 2.) Adjustment for nonresponse to the screener; 3.) Development of a combination weight to allow the black and Hispanic cases from the cross-sectional sample to be merged with those from the supplemental sample (non-Hispanic, non-blacks in the supplemental sample were not eligible for the NLSY97 sample); 4.) Adjustment of the weights for nonresponse to NLSY97 interviews; 5.) Poststratification of the nonresponse-adjusted weights to match national totals.

3. The National Longitudinal Study of Adolescent Health (Add Health), Carolina Population Center, University of North Carolina at Chapel Hill

Description: Add Health is the largest, most comprehensive longitudinal survey of adolescents ever undertaken. Beginning with an in-school questionnaire administered to a nationally representative sample of students in grades 7-12, the study followed up with a series of in-home interviews conducted in 1994-95, 1996, 2001-02, and 2007-08. Other sources of data include questionnaires for parents, siblings, fellow students, and school administrators and interviews with romantic partners. Preexisting databases provide information about neighborhoods and communities. A sample of 80 high schools and 52 middle schools from the US was selected with unequal probability of selection. Incorporating systematic sampling methods and implicit stratification into the Add Health study design ensured this sample is representative of US schools with respect to region of country, urbanicity, school size, school type, and ethnicity.

Study Content: chronic and disabling conditions, injury, mental health status (focus on depression), suicidal intentions/thoughts, health-service access and use, height, weight, blood pressure (available 2009), biomarkers for metabolic, immune and inflammatory processes (available 2009), medications (available 2009), personality, religiosity and spirituality, sleep patterns, physical activity, diet, substance use/abuse, violence, delinquency, criminal offending, and involvement with the criminal justice system, education history and high school transcripts, work experiences and military service, relationships (parents, teachers, friends, romantic partners), sexual behavior, sexually transmitted infections, contraception, pregnancy, children and parenting, household composition, neighborhood and school characteristics

Public Use Datasets: Public use datasets are available on their website (<http://www.cpc.unc.edu/projects/addhealth/data/publicdata>), along with detailed guides on how to use the data (<http://www.cpc.unc.edu/projects/addhealth/data/guides>) and program code.

Weights: Individual, paired (friends, couples, and siblings), and school level weights are available in the Add Health datasets for participants. The individual sampling weights are constructed using a nine step process that takes into account the probability of high school selection, school non-response and “feeder school” (middle school whose students attend the selected high school) selection probability, as well as student non-response and an adjustment to current population estimates

(<http://www.cpc.unc.edu/projects/addhealth/data/guides/weights.pdf>).

Using the sampling weights along with a variable to identify clustering of adolescents within schools, unbiased population parameters and standard errors can be obtained for single level analysis (<http://www.cpc.unc.edu/projects/addhealth/data/guides/wt-guidelines.pdf>). For multilevel models a scalable estimation technique and corresponding SAS program is provided: (http://www.cpc.unc.edu/research/tools/data_analysis/ml_sampling_weights).

Paired weights are constructed using variables of self-reported friends, romantic partners and siblings within the datasets, along with the individual weights and school weights. An additional trimming procedure was used to limit the value of the paired weights, and minimize the variance and bias of estimates (<http://www.cpc.unc.edu/projects/addhealth/data/guides/pweights.pdf>).