**Data Sources Description**

**NLSY79**

**Paragraph Description**

The NLSY79 Cohort is a longitudinal project that follows the lives of a sample of American youth born in the years 1957-1964. The individuals were interviewed for the first time in 1979, when they had between 14 to 22 years old. Two subsamples were dropped and 9,964 respondent remained. The data are now available for Round 1 (1979 survey year) and to Round 24 (2010 survey year).

**The three independent samples that compose the NLSY79**

Three independent probability samples compose the NLSY79.  These samples are designed to represent the entire population of youth aged 14 to 21 as of December 31, 1978, residing in the United States on January 1, 1979.  The three samples are:

1. A cross-sectional sample (6,111) designed to represent the noni-nstitutionalized civilian segment of young people living in the United States in 1979 and born January 1, 1957, through December 31, 1964
2. A set of supplemental samples (5,295) designed to oversample civilian Hispanic or Latino, black, and economically disadvantaged, nonblack/non-Hispanic youths born in the same time period
3. A military sample (1,280) designed to represent the population born January 1, 1957, through December 31, 1961, serving in the military as of September 30, 1978.  The inclusion of the military sample allows comparative civilian-military analyses.

**Based on this, I think we only have interest on the supplemental samples for blacks. In case we want to compare the sampling processes, I also include the sampling process for civilians.**

**Sampling Processes**

1. **Cross-Sectional Sample.**Following the initial screening process, 6,812 individuals from the cross-sectional sample were designated to be interviewed in the base year; of those, 90 percent or 6,111 respondents were actually interviewed in 1979.  The cross-sectional sample is designed to maximize the statistical efficiency of samples which are "cross-sectional" with respect to the rural population.  Specifically, through the several stages of sample selection (counties, enumeration districts-block groups, sample listing units), probabilities of selection are based upon either total population or total housing units.  Except for the economically disadvantaged supplemental sample, sampling of nonblack/non-Hispanic respondents was restricted to the 102 PSU National Sample.
2. **Supplemental Sample.**After screening, 5,969 individuals from the supplemental sample were designated for base year interviews, and of these, 89 percent or 5,295 respondents were actually interviewed.  Stratification specifically relevant for Hispanics or Latinos, non-Hispanic blacks, and economically disadvantaged, nonblack/non-Hispanics was used.  Probability proportional to size procedures were based on size measures for these groups rather than for the general population, making it possible to more nearly equalize the distribution of the targeted groups among the various sampling units than would otherwise be the case.

**What are the weights in the NLSY79 and how should we use them in our estimation?**

In each survey year a set of sampling weights is constructed.  These weights provide the researcher with an estimate of how many individuals in the United States each respondent's answers represent.

**Simple Example:**

**If we want to sample characteristics** for a single interview year in order to describe the population represented (that is, compute sample means, totals, or proportions), we need to weight the observations with the weights provided.  For example, to estimate the average hours worked in 1987 by persons born in 1957 through 1964, simply use the weighted average of hours worked, where weight is the 1987 sample weight.

**What are the available outcomes?**

As described in the .xls that I attach, and where I think it is a good idea to capture the available outcomes for all our data sources, we have: years of education; graduated from high school (does not include GED), not working or studying, married or in a serious relationship; had a child before turning 18 years old, yearly labor earnings (and we can put It in whatever terms we want).

**CNLSY79 (Pietro Birolo graciously provided me information on this data)**

The CNLSY79 considers the children of the people surveyed in the NLSY79. Its collection started in 1986 and ended up in 2010. The children were assessed in a bi-annual basis and put together detailed information on demographics, health, cognitive and non-cognitive abilities, etc. The data were collected directly via a the home and indirectly through mother reports.

**Some details are the following:**

* The child sample includes children born to female NLSY79 respondents. The child sample began in 1986, and the expanded mother-child data collection has occurred biennially since then.
* The number of children born to interviewed mothers has increased from 5,255 in 1986 to 8,099 in 2010.
* Interviews were completed during 2010 with 6,997 children (including 6,102 young adult children), or 86.4 percent of children born to mothers interviewed in that survey round.
* Starting with the 1994 survey, the children are treated as two separate groups.
* The first includes children who were under age 15 (as of December 31 of the survey year).  These children completed one or more of the assessment instruments, and information about each child is obtained from the child's mother.
* The second group comprises NLSY79 children who are at least 15 years of age by the end of the survey's calendar year.  These "Young Adults," most of whom were assessed during earlier child surveys, are administered a separate NLSY79-style questionnaire that gathers information on a wide range of topics

**Importantly: I talked to Pietro Biroli and he told me that, according to his experience, this sample is never used as representative. People in the NLSY79 had children in whatever patterns they had did and there are no sign that enable to claim representativeness of the sample at some level. I don’t think this is a good data source for comparison by itself.**