Add Health Data Documentation

Siying Cao [*]

April 8, 2017

1 Introduction to Add Health Survey

The National Longitudinal Study of Adolescent to Adult Health (Add Health) survey is a nationally representative study that explores the health-related behaviors of adolescents in grades 7 through 12. The first wave of survey was conducted during 1994-1995, and sub-samples of the subjects have been followed through adolescence and the transition to adulthood with four in-home interviews. In particular, the survey seeks to examine how social contexts (families, friends, peers, schools, neighborhoods, and communities) influence adolescents health and risk behaviors. With such an objective, the Add Health survey includes detailed information about respondents demographics, family background, academic outcomes, and health-related behaviors, as well as their social networks. More information can be found at the official website `http://www.cpc.unc.edu/projects/addhealth`

2 Data Sets Description

In the data repository, you will find complete public-use datasets spanning all four waves.

Wave I contains four different data sets for a total of 6,504 individual observations, the core of which is the **In-Home** Data File. There are 2,801 variables covering a wide range of topics, including the respondents demographics, family background, health-related behaviors, and sexual relationship. In the **Contextual** Data File, you will find 32 variables on community population and housing characteristics at block group levels [1] derived from the Wave I addresses. The **Network** Data File contains 439 variables constructed from the in-school data and friendship nominations, in which the respondent was asked to nominate up to five male and five female friends. The public-use version dataset does not contain friend identification numbers, making it impossible to link a respondents

[*] Please send any comments to `siyingc -at- uchicago -dot- edu`

[1] Block group is a census defined geographic area that averaged 452 housing units, or 1,100 people. It is the lowest level of geographic unit for which Census Bureau publishes sample data

information to her friends. However, the network data provides plenty of interesting information on network structure that you may find useful to exploit.

Wave II interviewed 4,834 of the original Wave I respondents in 1996 and the questions were generally similar to that at Wave I. **In-Home** (2,532 variables) and **Contextual** (32 variables) Data Files are provided. Wave III data were collected during 2001-2002, in which 4,882 of the Wave I respondents were surveyed. As adolescents transition to young adulthood, additional measures on educational outcome, relationship, and fertility are constructed in **In-Home (multiple files), Education, and Graduation** Data File. Finally Wave IV collected data on 5,114 respondents of the original Wave I sample, aged 24-32 to study the trajectories to adulthood. Apart from the usual **In-home** Data File, biological data was also gathered to understand pre-disease pathways, with a special focus on obesity, stress, and health risk behavior. Such data are reported in **Biomarker** File.

Data are downloaded from the Inter-University Consortium for Political and Social Research (ICPSR)[2] and are well formatted to both STATA and R users. Each data set also includes a codebook/questionnaire to help you understand the survey questions and variables associated. You can ignore the frequency table file coming with it.

3 Research Guide

Given the nature of this survey, both cross-sectional and longitudinal type of research are possible. Since individual identifier (`AID`) appears as the leading variable in all data files, data from multiple files across waves can be easily merged to study social interaction potentially in many interesting ways.

The Add Health study design used a clustered sample in which the clusters were sampled with unequal probability. As a result, to analyze the data correctly requires the use of proper weights with the aid of survey software packages designed for handling such complication.

Weight variables are supplied in each wave as a standalone data file, and can be directly merged to the main data set to perform weight-adjusted analysis. Table 2.4 and 2.5 of the User Guide [3] provides a useful summary of which weight variables to use depending on the wave(s) you are leveraging in either cross-section or panel analysis.

[2] http://www.icpsr.umich.edu/icpsrweb/ICPSR/studies/21600?archive=ICPSR&q=21600

[3] https://github.com/ECON31830/AddHealthData/blob/master/User_guide.pdf