Guidelines for Analyzing Add Health Data

Ping Chen

Kim Chantala

Carolina Population Center
University of North Carolina at Chapel Hill

Last Update: March 2014

https://doi.org/10.17615/C6BW8W

Table of Contents

Overview

The National Longitudinal Study of Adolescent Health (Add Health) is a longitudinal study of a nationally representative sample of adolescents in grades 7-12 in the United States in 1994-95 that has been followed through adolescence and the transition to adulthood with four in-home interviews. The Add Health study design used a clustered sample in which the clusters were sampled with unequal probability. While reducing the cost of data collection, this design complicates the statistical analysis because the observations are no longer independent and identically distributed. To analyze the data correctly requires the use of special survey software packages specifically designed to handle observations that are not independent and identically distributed. The purpose of this document is to provide guidelines to correctly analyze the Add Health data. To do this, we describe the characteristics of the Add Health sample design and data elements needed by the survey software packages. We next identify a series of common errors to avoid when analyzing the Add Health data. Lastly, we provide examples of different types of analysis using various survey software packages.

Chapter 1. Basic Concepts of the Add Health Design

This section describes how the Add Health sample was selected and discusses attributes of the Add Health sample that can impact analysis.

Understanding the Add Health Sampling Design

Add Health is a longitudinal study of adolescents enrolled in 7th through 12th grade in the 1994 - 1995 academic year. Add Health used a school-based design. The primary sampling frame was derived from the Quality Education Database (QED) comprised of 26,666 U. S. High Schools. From this frame we selected a stratified sample of 80 high schools (defined as schools with an 11th grade and more than 30 students) with probability of selection proportional to school size. Schools were stratified by region, urbanicity, school type (public, private, parochial), ethnic mix, and size. For each high school selected, we identified and recruited one of its feeder schools (typically a middle school) with probability proportional to its student contribution to the high school, yielding one school pair in each of 80 different communities. More than 70 percent of the originally selected schools agreed to participate in the study. Replacement schools were selected within each stratum until an eligible school or school-pair was found. Overall, 79 percent of the schools that we contacted agreed to participate in the study. A total of 52 feeder (junior high & middle) schools were selected. Because some schools spanned grades 7 to 12, we have 132 schools in our sample, each associated with one of 80 communities. School size varied from fewer than 100 students to more than 3,000 students. Our communities were located in urban, suburban, and rural areas of the country. Administrators at each school were asked to fill out a special survey that captured attributes of the school.

Add Health has collected multiple waves of data on adolescents recruited from these schools, as follows:

In-School Survey (1994): Over 90,000 students completed a questionnaire. Each school administration occurred on a single day within one 45- to 60-minute class period.

Wave I In-Home Survey (1995): Adolescents were selected with unequal probability of selection from the 1994-1995 enrollment rosters for the schools and those not on rosters that completed the in-school questionnaire. A core sample was derived from this administration by stratifying students in each school by grade and sex and then randomly choosing about 17 students from each stratum to yield a total of approximately 200 adolescents from each pair of schools. The core in-home sample is essentially self-weighting, and provides a nationally representative sample of 12,105 adolescents in grades 7 to 12. Further, we drew supplemental samples based on ethnicity (Cuban, Puerto Rican, and Chinese), genetic relatedness to siblings (twins, full sibs, half sibs, and unrelated adolescents living in the same household), adoption status, and disability. We also oversampled black adolescents with highly educated parents.

Wave II In-Home Survey (1996): Participants from Wave I excluding adolescents in 12th grade at Wave I interview who were not part of the genetic sample[1]. Some adolescents not interviewed at Wave I were interviewed at Wave II in order to increase the number of respondents in the genetic sample.

Wave III In-Home Survey (2001): Participants from Wave I In-home Survey. Participants interviewed only at Wave II were also included if they were part of the genetic sample. 687 cases from Wave I, without sampling weights and not in the genetic sample, were not included.

Wave IV In-Home Survey (2008): Participants from Wave I In-home Survey. The 687 cases not sampled at Wave III were also excluded at Wave IV.

A detailed list of attributes for selecting schools and adolescents appears in Table1.1. All attributes listed in Table1.1, as well as characteristics related to non-response, were employed to compute the final sampling weights. For each panel of data collection, Add Health provides sampling weights that are designed for estimating single-level (population-average) and multilevel models. These weights are available for both schools and adolescents. For additional details about the Add Health sampling design, see Harris (2013) and Tourangeau and Shin (1999).

Impact of the Sampling Design on Analysis

Unless appropriate adjustments are made for sample selection and participation, estimates from analyses using the Add Health data can be biased when any factor used as a basis for selection as a participant in the Add Health Study also influences the outcome of interest. For example, black adolescents whose parents were college graduates comprise one of the many over-sampled groups. Parental education is a factor that affected selection of black youth in the Add Health study and can also influence family income. Unless the analytic technique uses appropriate statistical methods to adjust for over-sampling, estimates of the income of blacks will be biased. Any analysis that includes family income or other variables related to family income may produce biased estimates unless proper adjustments are made for over-sampling.

To obtain unbiased estimates, it is important to account for the sampling design by using analytical methods designed to handle clustered data collected from respondents with unequal probability of selection. Failure to account for the sampling design usually leads to under-estimating standard errors and false-positive statistical test results. Table 1.2 lists the attributes of the Add Health sampling design that should be taken into consideration during analysis.

[1] The genetic sample consists of pairs of siblings living in the same households, identical twins, fraternal twins, full siblings, and half siblings in addition to non-related pairs, such as step-siblings, foster children, and adopted (non-related) siblings.

Table 1.1. Attributes of the Add Health sampling design influencing selection of adolescents for recruitment

	Sampled Unit	
	Schools	Adolescents
Attributes related to being selected to participate in Add Health	*HIGH SCHOOLS:* *Size of School:* <125 students 126-350 students 351-775 students ≥776 students *School Type:* public private parochial *Location:* urban suburban rural *Region:* Northeast Midwest South West *Percent White:* 0 % 1 to 66 % 67 to 93% 94 to 100% *FEEDER SCHOOLS:* Percent of entering class for linked High School coming from the feeder school *	*WAVE I ADOLESCENTS:* *Race/Ethnicity over-sampled Groups:* High SES Black Cuban Puerto Rican Chinese *Genetic Sample* Twins Full Siblings Half Siblings Unrelated in Same Household *Disabled Youth over-sampled Group* *Purposively Selected Schools:* All students selected from 16 schools
Panels of Data affected by Attribute of Sampled Unit	*School Administrator* *In-School* *Wave I* *Wave II* *Wave III* *Wave IV*	*Wave I* *Wave II* *Wave III* *Wave IV*

Table 1.2. Attributes of Add Health Sampling Design

Design Attribute	Usual Impact on Analysis	Variables in Add Health Data Used to Adjust for the Sampling Design
Stratification	Reduce Variance	POSTSTRATIFICATION VARIABLE: Census Region
Clustering of Students	Increase Variance	PRIMARY SAMPLING UNIT VARIABLE: School Identification Variable
Unequal Probability of Selection	Increase Variance	SAMPLING WEIGHTS: • Cross-sectional Weights for Schools • Cross-sectional Weights for analyzing each Wave of Data • Cross-sectional Weights for analyzing special sub-samples from Wave III • Longitudinal Weights for conducting analyses combining data from multiple Waves • Multilevel Weights for two-level analysis where schools and adolescents are the levels of interest

Chapter 2. Choosing the Correct Sampling Weight for Analysis

The Add Health sampling weights are designed to turn the sample of adolescents we interviewed into the population we want to study. These weights are available for the respondents who are members of the Add Health probability sample. By using these sampling weights and a variable to identify clustering of adolescents within schools, you can obtain unbiased estimates of population parameters and standard errors from your analysis. This chapter describes the sampling weights distributed with the Add Health data and provides instruction on which weight should be used in your analysis.

Available Sampling Weights

The Add Health sampling weights were developed for analyzing combinations of data from the In-Home Interviews using a variety of techniques. Usage of these weights can be divided into three different categories of analyses.

Single-Level (Population-Average) Model

The first category includes analyses to provide population estimates for adolescents who were enrolled in school for the 1994-1995 academic year (see Table 2.1). Often these analyses involve fitting a population-average (single-level or marginal) model. In Add Health, users usually use individual (respondent)-level data to estimate models.

Multilevel Model

The second category includes analyses fitting a multilevel model to provide estimates for adolescents who were in school during the 1994-1995 academic year. These weights are designed to estimate a model where the levels of interest in the analysis match the sampling levels of school and adolescent (Table 2.2). A weight component is available for each level of sampling (schools and adolescents) at each wave of data. These weight components differ in meaning from the sampling weights designed for estimating population-average (single-level) models that have been traditionally distributed with the Add Health data. They are used for analysis that includes both school-level and individual level data. They are the basic building blocks needed for computing the multilevel weights with the methods detailed in Chantala et al (2011). Be sure to scale the weight components listed in Table 2.2 by using the methods discussed in the above-linked document. Note that there is no weight component variable for neighborhood-level data because Add Health does not include neighborhood in its sampling design.

In a single level model, only a single grand sample weight is needed. The grand sample weight reflects the inverse of the probability of ultimate selection; here, "ultimate" means that it factors in all levels of clustered sampling, corrections for nonresponse, oversampling, and post-stratification, etc. In a single-level model, the use of the grand sample weight, w_{ij} is sufficient; w_{ij} is an unconditional weight for observation i, j.

In a two-level model with Add Health data, it is not sufficient to use the single grand sampling weight w_{ij}, because weights enter into the log likelihood at both the school level and individual level. Instead, required for a two-level model under this sampling design is w_j (the inverse of the probability that school *j* is selected in the first stage), and $w_{i|j}$ (the inverse of the probability that individual *i* from school *j* is selected at the second stage conditional on school *j* already being selected). It is not appropriate in this case to use only the grand sample weight w_{ij} without making assumptions about w_j.

Table 2.1. Sampling Weights distributed with the Add health data designed for estimating single-level (marginal or population average) models.

Data Set (Year collected)	Sampling Weight Variable (N)	Type	Sample	Target Population
Wave I (1995)	GSWGT1 (N=18,924)	Cross-sectional weight	Adolescents chosen with a known probability of being selected from 1994-1995 enrollment rosters of US schools	Grade 7-12[1] in 1994-1995
Wave II (1996)	GSWGT2 (N=13,570)	Cross-sectional weight	Adolescents interviewed at Wave II. 13,568 of these adolescents were also interviewed at Wave I	Grade 7-11[1] in 1994-1995
Wave III (2001)	GSWGT3_2 (N=14,322)	Cross-sectional weight	Wave I respondents who were interviewed at Wave III	Grade 7-12[1] in 1994-1995
Wave III (2001)	GSWGT3 (N=10,828)	Longitudinal weight	Eligible Wave I Respondents interviewed at both Wave II & Wave III	Grade 7-11[1] in 1994-1995
Wave IV (2008)	GSWGT4_2 (N=14,800)	Cross-sectional weight	Wave I respondents who were interviewed at Wave IV	Grade 7-12[1] in 1994-1995
Wave IV (2008)	GSWGT4 (N=9,421)	Longitudinal weight	Eligible Wave I respondents who were interviewed at Wave II, III & IV	Grade 7-11[1] in 1994-1995
Wave IV (2008)	GSWGT134 (N=12,288)	Longitudinal weight	Eligible Wave I respondents who were interviewed at Wave III & IV	Grade 7-12[1] in 1994-1995

[1] The Target Population for these samples is comprised of adolescents who were enrolled in US schools during the 1994-1995 academic year for the specified grades.

Table 2.2. Available In-Home Weight Components for Multilevel Analyses involving the In-School, Wave I, II, III and IV data sets.

Interview (Year collected)	Level 2 Weight Component (N)	Level 1 Weight Component (N)	Sample	Target Population	Type
In-School (1994)	SCHWT128 (N=128)	INSCH_WT (N=83,135)	Adolescents chosen with a known probability of being selected from 1994-1995 enrollment rosters of US schools	Grade 7-12 in 1994-1995	Cross-sectional weights
Wave I (1995)	SCHWT1 (N=132)	W1_WC (N=18,924)	Adolescents chosen with a known probability of being selected from 1994-1995 enrollment rosters of US schools	Grade 7-12 in 1994-1995	Cross-sectional weights
Wave II (1996)	SCHWT1 (N=132)	W2_WC (N=13,568)	Adolescents interviewed at Wave II. 13,568 of these adolescents were also interviewed at Wave I	Grade 7-11 in 1994-1995	Cross-sectional weights
Wave III (2001)	SCHWT1 (N=132)	W3_2_WC (N=14,322)	Wave I respondents who were interviewed at Wave III	Grade 7-12 in 1994-1995	Cross-sectional weights
Wave III (2001)	SCHWT1 (N=132)	W3_WC (N=10,828)	Eligible Wave I Respondents interviewed at both Wave II & Wave III	Grade 7-11 in 1994-1995	Longitudinal weight
Wave IV (2008)	SCHWT1 (N=132)	W4_2_WC (N=14,800)	Wave I respondents who were interviewed at Wave IV.	Grade 7-12 in 1994-1995	Cross-sectional weights
Wave IV (2008)	SCHWT1 (N=132)	W4_WC (N=9,421)	Eligible Wave I Respondents interviewed at Wave II, III & Wave IV	Grade 7-11 in 1994-1995	Longitudinal weight

Both the school-level w_j and individual-level $w_{i|j}$ are called weight components in Add Health. As mentioned earlier, if both the school-level and individual-level weight components are included in the two-level model, rescaling is necessary to remove the dependence of $w_{i|j}$ on w_j. Further details on weighting and scaling in xtmixed with Survey data are available in the Stata manual (p. 342-343).

Single-Level Model for Special Subpopulation

The third category includes analyses fitting a population-average model for special subpopulations in the US who were enrolled in school for the 1994-1995 academic year (Table 2.3). Special sub-samples of the Wave III respondents were selected for additional testing or special sections of the Wave III survey.

The *Romantic Partner* sample is comprised of 1,317 Wave III respondents and their romantic partners. This sample was selected at Wave III to study relationship commitment and intimacy. The recruitment criteria were:

- Current romantic relationship
- Heterosexual relationship
- Partner and Add Health respondent are at least 18 years old
- Relationship has lasted at least 3 months

Approximately equal numbers of married, cohabiting, and dating couples were recruited into the study. The entire Wave III questionnaire was completed by both the Add Health respondent and their partner.

The *Wave III Educational Sample* is comprised of the Wave III respondents whose high school transcripts were available for collection. Transcript availability was affected by many issues unrelated to the nonresponse adjustments made to the Wave III grand sample weights. For example, transcripts were unavailable if the Wave III respondent did not attend high school, was home-schooled, or attended school outside of the US. In addition, transcripts were not collected if the school was closed, refused to provide students' transcripts, or provided incomplete or incorrect transcripts. Because of this, special sampling weights were constructed to adjust for transcript nonresponse as well as survey nonresponse. Using these sampling weights in analyses that incorporate transcript information will reduce bias in estimates and standard errors.

The *MGEN* original sample included 2,932 (cross-sectional) and 2,195 (longitudinal) Wave III male and female respondents, who were randomly flagged to have their urine assayed for *mycoplasma genitalium*. A number of post-stratification variables were selected to calibrate the weights of the 2,932 assayed cases to all 14,322 respondents in the cross-sectional sample and the 2,195 assayed cases to all 10,828 respondents in the longitudinal sample.

Table 2.3. Sampling Weights distributed with the Add Health data designed for estimating single-level (marginal or population average) models.

Data Set (Year collected)	Sampling Weight Variable (N)	Sample	Target Population
Wave III (2001)	W3PTNR (N=1,317)	Wave III Romantic Partner Sample: Eligible Wave I respondents and romantic partners interviewed at Wave III	Romantic Partners [2]
	TWGT3_2 (N=11,637)	Wave III Education Sample: Eligible Wave I respondents interviewed at Wave III	Grade 7-12[1] in 1994-1995
	TWGT3 (N=8,847)	Wave III Education Sample: Eligible Wave I respondents interviewed at Wave II and III	Grade 7-11[1] in 1994-1995
	MGENCRWT (N=14,322) (MGEN Cross-Sectional Weight)	MGEN Sample: special sample selected for testing urine for mycoplasma genitalium at Wave III	Grade 7-12[1] in 1994-1995
	MGENLOWT (N=10,828) (MGEN Longitudinal Weight)	MGEN Sample: special sample selected for testing urine for mycoplasma genitalium. Eligible Wave I respondents interviewed at Wave II and III	Grade 7-11[1] in 1994-1995
	HPVCRWT (N=6,593) (HPV Cross-Sectional Weight)	HPV Sample: special sample of sexually active females selected for testing urine for Human Papillomavirus at Wave III	Sexually Active Female Population
	HPLORWT (N=4,945) (HPV Longitudinal Weight)	HPV Sample: special sample of sexually active females selected for testing urine for Human Papillomavirus. Corresponding Wave I respondents interviewed at Wave II and III	Sexually Active Female Population

[1] The Target Population for these samples is comprised of adolescents who were enrolled in US schools during the 1994-1995 academic year for the specified grades.

[2] The Target Population for the Wave III Romantic Partner Sample is Couples in 2001 where at least one member of the couple was enrolled in US schools during the 1994-1995 academic year for the specified grades.

The *HPV* original sample included 3,369 (cross-sectional) and 2,535 (longitudinal) Wave III sexually active female respondents, who were randomly flagged to have their urine assayed for *human papillomavirus*. A number of post-stratification variables were selected to calibrate the weights of the 3,369 assayed cases to all 6,593 sexually active female respondents in the cross-sectional sample and the 2,535 assayed cases to all 4,945 sexually active female respondents in the longitudinal sample. Detailed documentation for the HPV and MGEN weights are provided with the restricted-use data for these results or by request from addhealth@unc.edu.

Choosing a Sampling Weight for Analysis

The sampling weight selected for an analysis depends on both the type of analysis required to investigate a hypothesis and the interview or combination of interviews needed in the analysis. The following section gives instructions on selecting the best sampling weight for different types of analysis.

Cross-Sectional Analysis

Research questions addressed by cross-sectional analysis are those that investigate association rather than causation. The temporal sequence of events necessary for drawing causal inferences may not be available. Data for both predictive and outcome variables are collected at the same point in time, that is, from the same wave. The outcome can be observed for all subjects. The correct choice of sampling weight in this instance would be the weight that was created for everyone in the probability sample for the wave of data used (Table 2.4).

Another scenario is when the outcome variable is from one wave of data, i.e., Wave I, II, III or IV, but the predictors (or covariates) are from either previous wave(s) or a combination of waves. Under these circumstances, the correct weight would be the cross-sectional weight for the wave from where the outcome variable comes, rather than the longitudinal weight (see Table 2.4). If you are using data from multiple waves for covariates (predictor variables), you might also need to use the subpopulation option (see example 3 in Chapter 4).

Longitudinal Analysis

Longitudinal analysis is used to address research questions that investigate changes in measurements taken on the same respondents over time, that is, the outcome variable is measured multiple times. *Note that if the covariates are from multiple waves but the outcome variable is from just one wave of data, you do not need to use the longitudinal weight.*

The outcome can be observed for all subjects and the data being analyzed can be organized in different ways. Two common ways are:

- one record per respondent (AID) per time point
- multiple records for a respondent can be combined so that each new record is constructed by computing the difference in values of variables collected at each point in time.

A potential difficulty in longitudinal analysis is that the measurements for a respondent may be missing at one or more time points. Sampling weights incorporating a non-response adjustment have been created to compensate for data missing at a particular time point because the respondent was not interviewed. The analyst only then need consider the effect of item non-response rather than both item and survey non-response.

Longitudinal analysis with the Add Health data will use information collected from interviews at two or more time-points (waves) for the outcome variable. In general, the choice of sampling weight for longitudinal analysis will be determined by the data collected at the most recent time-point. Table 2.5 shows the appropriate sampling weight to use for most longitudinal analyses that estimate population-average models.

Time-to-Event Analysis

Research questions best answered by time-to-event analysis are those involving the occurrence and timing of events. Data comes from individuals observed over time where the outcome is the occurrence of a specific event that is a qualitative change that can be situated in time. Large and sudden changes in quantitative variables can also be treated as events. Example events are death, onset of disease, first pregnancy, or loss of virginity. The event is not observed for all respondents. Choice of sampling weight will usually be determined by the data collected at the earliest time point.

Summary

The guidelines presented in this chapter for choosing the correct sampling weight for most analyses can be summarized in three simple rules:

1. Cross-Sectional Analysis: Choose the weight created for everyone in the probability sample (see Table 2.4) for the population of interest.

2. Longitudinal Analysis: Choose the weight from the Wave of data collected at the latest time-point (see Table 2.5) for the population of interest.

3. Time-to-Event Analysis: Choose the weight from the Wave of data collected at the earliest time point (see Table 2.6) for the population of interest.

These rules should allow the analyst to select the best sampling weight for most research endeavors.

Table 2.4. Sampling Weights used in Cross-sectional Analysis

Population of Interest	Data Used	Number of Participants in Analysis File	Sampling Weight Population Average Models	Sampling Weight Multilevel Models
Adolescents in 1995 enrolled in Grade7-12 during 1994-1995	Wave I	18,924	GSWGT1	SCHWT1 W1_WC
Adolescents in 1996 enrolled in Grade 7-11 during 1994-1995	Wave II	13,570	GSWGT2	SCHWT1 W2_WC
Young Adults in 2001 enrolled in Grade 7-12 during 1994-1995	Wave III	14,322	GSWGT3_2	SCHWT1 W3_2_WC
Young adults Romantic Couples in 2001 (one partner enrolled in Grade 7-12 during 1994-1995)	Wave III	1,317	W3PTNR	Not Available
Young Adults in 2001 enrolled in Grade 7-12 during 1994-1995 (Educational analyses involving high school transcripts)	Wave III	11,637	TWGT3_2	Not Available
Young Adults in 2001 enrolled in Grade 7-12 during 1994-1995 (Analyses involving special sample selected for testing urine for mycoplasma genitalium at Wave III.)	Wave III	14,322	MGENCRWT	Not Available
Sexually Active Female Population	Wave III	6,593	HPVCRWT	Not Available
Young Adults in 2008 enrolled in Grade 7-12 during 1994-1995	Wave IV	14,800	GSWGT4_2	SCHWT1 W4_2_WC

Table 2.5. Sampling Weights used for Longitudinal Analysis

Population of Interest is Represented By	Data Used	Number of Subjects in Analysis File	Sampling Weight for Population Average Models	Sampling Weight for Multilevel Models
Adolescents enrolled in Grade 7-11 during 1994-1995 interviewed in 1995 & 1996	Wave I & II	13,568	GSWGT2	SCHWT1 W2_WC
Adolescents enrolled in Grade 7-12 during 1994-1995 interviewed in 1995 & 2001	Wave I & III	14,322	GSWGT3_2	SCHWT1 W3_2_WC
Adolescents enrolled in Grade 7-11 during 1994-1995 interviewed in 1996 & 2001	Wave II & III	10,828	GSWGT3	SCHWT1 W3_WC
Adolescents enrolled in Grade 7-11 during 1994-1995 interviewed in 1995, 1996 & 2001	Wave I, II, & III	10,828	GSWGT3	SCHWT1 W3_WC
Adolescents enrolled in Grade 7-11 during 1994-1995 interviewed in 1996 & 2001 (Educational analyses involving high school transcripts)	Wave II & III	8,847	TWGT3 (N=8,847)	Not Available
Adolescents enrolled in Grade 7-11 during 1994-1995 and interviewed in 1995, 1996, & 2001 (Analyses involving MGEN sample)	Wave I, II, III	10,828	MGENLOWT	Not Available
Sexually Active Female Population (Analyses involving HPV sample)	Wave I, II, III	4,945	HPLORWT	Not Available
Adolescents enrolled in Grade 7-11 during 1994-1995 interviewed in 1995, 1996, 2001 & 2008	Wave I, II, III & IV	9,421	GSWGT4	SCHWT1 W4_WC
Adolescents enrolled in Grade 7-12uring 1994-1995 interviewed in 1995, 2001 & 2008	Wave I, III & IV	12,288	GSWGT134	Not Available

Table 2.6. Sampling Weights used for Time-to Event Analysis

Data availability and Population of Interest is Represented by	Data Source	Number in Analysis File	Weight for Population Average Models	Weights for Multilevel Models
Data available from only one interview:				
Adolescents in 1995 enrolled in Grade7-12 during 1994-1995	Wave I only	18,924	GSWGT1	SCHWT1 W1_WC
Adolescents in 1996 enrolled in Grade 7-11 during 1994-1995	Wave II only	13,570	GSWGT2	SCHWT1 W2_WC
Young Adults in 2001 enrolled in Grade 7-12 during 1994-1995	Wave III only	14,322	GSWGT3_2	SCHWT1 W3_2_WC
Young Adults in 2008 enrolled in Grade 7-12 during 1994-1995	Wave IV only	14,800	GSWGT4_2	SCHWT1 W4_2_WC
Data available from Multiple interviews:				
Adolescents in 1995 enrolled in Grade7-12 during 1994-1995	Wave I & II	18,924	GSWGT1	SCHWT1 W1_WC
Adolescents in 1996 enrolled in Grade 7-11 during 1994-1995	Wave II & III	13,570	GSWGT2	SCHWT1 W2_WC
Young Adults in 2001 enrolled in Grade 7-12 during 1994-1995	Wave I, II, & III	14,322	GSWGT1	SCHWT1 W1_WC
Young Adults in 2008 enrolled in Grade 7-12 during 1994-1995	Wave I, II, III & IV	14,800	GSWGT1	SCHWT1 W1_WC

Analyzing Pairs of Respondents

Some analyses of interest will involve serendipitous pairs of respondents. Such pairs may be comprised of unrelated friends, twins, or other siblings. For example, the Add Health data includes respondents who are friends with each other. Thus, in your model, you may be predicting an outcome that uses survey responses from both the respondent and the friend. The choice of weights for analysis that includes observations based on data from two different but connected respondents is not straightforward. One acceptable method is to calculate the weight by first computing the joint inclusion probability of each pair, then deriving its inverse; this value will serve as the weight. In any circumstances where there are two related or connected respondents, it is essential to examine the details of the sample selection procedure for both of the individuals and their schools. The selection sample procedure may vary for each type of pair (i.e., friends, siblings, twins, and romantic partners), requiring a different method of computing the weight for the

specific type of pair. Add Health has constructed weights for the romantic partners sample (see Table 2.3). You can use these weights for partners analysis if you agree with the computational adjustments illustrated in the documentation. Otherwise, we suggest you consult a statistician before constructing special weights for any type of pairs analysis.

Genetic Sample Weights

Add Health Wave I data includes a genetic supplemental sample. The genetic sample was selected based on the sibling relationships in which the student was involved: (1) twins; any student who identified himself or herself as a twin was included in the twin supplement; (2) other siblings of twins; (3) other full siblings, including brother pairs, sister pairs, and brother-sister pairs; (4) half siblings, where both members of the pair were enrolled in grades 7 through 12; and (5) unrelated (adolescents enrolled in grades 7 through 12 who did not share a biological mother or father but who are living in the same household). Genetic sample weights are not needed when using any data from this genetic supplemental sample.

Add Health has two types of weights available for use with the genetic sample, one for analyses when the analysis unit is individuals, the other for analyses when the analysis unit is pairs. Variables derived from household information (that is, the household from where the pair comes from, rather than the individual adolescents in the pair), including race, educational level, and marital status, are used as post-stratification variables. The 1995 Current Population Survey was selected as the calibration population. Weights for the genetic sample and corresponding documentation are available from Add Health (addhealth@unc.edu). Researchers using these weights should have a good understanding of and be in agreement with the weighting procedure, as subsequent results can be generalized only to a 1995 US population of persons or pairs of individuals, aged 12 to 18, who live in the same household. The biological relationships of these within-household persons/pairs are unknown. We suggest that researchers using these weights provide statistical results for analyses conducted both with and without weights for comparison.

Wave III Binge Sample

The binge sample includes participants selected at Wave III to study binge-drinking attitudes among college-age students. The eligibility criteria for inclusion in the binge sample were:

- In the 7th or 8th grade during Wave I
- Interviewed at both Wave I and II
- Never married at Wave III

At Wave III, questions 50—93, Section 28, were asked of approximately equal numbers of respondents in four groups who met eligibility criteria: females attending college, males attending college, females not attending college, and males not attending college. No weight variable is available for analyses using data from this sample. If you use the binge sample, be sure to read the documentation thoroughly, describe the sample in detail in all publications and presentations, and report that results from the binge sample cannot be generalized to the population.

Table 2.7. Sampling Weights for Wave I Genetic Sample with Single-Level Models

Data Set (Year collected)	Sampling Weight Variable (N)	Sample	Target Population
Wave I (1995)	PERSONWEIGHT (N=5,530)	Genetic sample of individuals with varying genetic resemblance, including monozygotic twins, dizygotic twins, full siblings, half siblings, and unrelated siblings who were raised in the same household.	1995 US population of individuals ages 12 to 18 who live in the same household.
	PAIRWEIGHT (N=3,160)	Genetic sample of pairs with varying genetic resemblance, including monozygotic twins, dizygotic twins, full siblings, half siblings, and unrelated siblings who were raised in the same household.	1995 US population of pairs of individuals ages 12 to 18 who live in the same household.

Chapter 3. Avoiding Common Errors

This chapter lists the most common errors made when analyzing Add Health data and how to avoid them. These recommendations focus on use of the probability sample to make estimates that are nationally representative. We conclude with a list of steps to take when preparing your data for analysis that will help avoid these errors.

3.1 Common Errors

Ignoring clustering and unequal probability of selection when analyzing the Add Health data

This results in biased estimates and false-positive hypothesis test results. Point estimates (means, regression parameters, proportions, etc.) are affected only by the weights. Variance estimates are affected by clustering, stratification, weights, and design type.

The easiest way to adjust estimates for clustering and unequal probability of selection is to use a survey software package that adjusts for clustering and uses sampling weights when computing point estimates and standard errors. This method is called *design-based analysis*. It is easy to implement and generate correct results because the design features, including design variables and error terms regarding the correlation structure of the data, are automatically incorporated by the survey software packages.

If the software package you are using does not allow you to specify sampling weights then you should include the covariates in your analysis that relate to the schools and adolescents being selected for participation in the Add Health Survey. These sampling attributes are listed in Table 1.1. (see Chapter 1). This method is called *model-based analysis*. However, it can be very difficult and time consuming to produce acceptable results with model-based analytic methods. You must understand how to incorporate detailed characteristics of the sampling plan, weighting scheme, and intra-cluster correlation (ICC), as well as the formulas used by the traditional statistical package and the adjustments that might need to be made to these formulas. We do not recommend this method unless you have previous experience using it.

In Table 3.1, we have classified analysis techniques into five different approaches. Ignoring both the weights and the design structure produces incorrect point estimates and variances. However, including weights in an analysis in which the design structure is ignored only gives correct point estimates (totals and ratios). If you only need point estimates and your standard software package allows you to use weights, there is no need to use other survey software packages. Note that using normalized weights produces incorrect estimates of the totals such as, for example, the total number of adolescents in the population.

Table 3.1. Comparison of Techniques Used to Analyze Survey Data

	Ignore Design Structure			Incorporate Design Structure	
	Model-Based Analysis			Model-Based Analysis	**Design-Based Analysis**
Effects on	Ignore Weights	Use Weights	Use Normalized Weights	Use Weights, Strata, Cluster	**Use Weights, Strata, Cluster**
Estimates of totals	Incorrect	Correct	Incorrect	Correct	**Correct**
Estimates of ratios, such as proportions, means, & regression parameters	Incorrect	Correct	Correct	Correct	**Correct**
Estimates of variances, standard errors, & confidence intervals	Incorrect	Incorrect	Incorrect	Close to correct	**Correct**

Including respondents who are missing sampling weights in analyses when your goal is to obtain national estimates. At Wave I, additional adolescents were selected outside of the sampling frame as part of the genetic sample. This was done to ensure that the sample size of genetically related individuals was large enough for specialized genetic analyses. Since these adolescents were selected outside of the sampling frame, sampling weights could not be constructed. Although the survey software will eliminate those adolescents who have a missing value for a sampling weight from the analyses, you may erroneously include them when determining the sample size.

***Subsetting the probability sample (i.e., adolescents who have weights) when using the survey software*.** When analyzing data from a sample survey, analyzing a subset of the sample is not the same as analyzing a subpopulation represented by part of the sample. For example, your interest may lie in performing an analysis on Asians only. Samples of students selected from some schools might not include any Asian students. However, if the sampling was repeated, some Asian students might be selected from these schools and the schools would remain in the analysis. The possible variation in school sample size that might occur in re-sampling must be included in estimating variances and standard errors. *Subsetting the data by deleting cases that are not in subsample may cause an incorrect number of PSU's to be used in the variance computation formula.* Most software packages for analyzing data from sample surveys provide special commands for using subpopulation analysis.

Using the Sampling Weight as a Frequency or Analytical Weight during Analysis. There are different types of weights used by the various software packages. The three most common types are:

Frequency Weights. These weights represent the number of respondents who were actually interviewed. For example, a frequency weight of 3 means that the three respondents were interviewed and all gave identical answers to a specific question.

Analytical or Variance Weights. These weights are inversely proportional to the variance of an observation. This type of weight might be used for data sets where the variables are averages across a group of individuals (or time points), where the weight is the number of elements used to compute the average.

Sampling Weights. These weights are computed as the inverse of the probability that a specific respondent was selected for the interview. A sampling plan will be used to guide the selection process of individuals to be recruited for participation in the survey. For example, a sampling weight of 25 means that the data from the recruited individual is representative of 25 respondents in the population of interest.

Each of these weights enters the computation in a different way and will give different estimates of variance and standard errors. Software packages do not always give different statements to uniquely define the type of weight. For example, the SAS statement:

WEIGHT GSWGT1;

will be used as a frequency weight in PROC FREQ, a variance weight in PROC REG, and a sampling weight in PROC SURVEYREG. On the other hand, Stata uses special keywords (*fweights* for frequency weights, *aweights* for analytical weights, and *pweights* for sampling weights) to specify how the weight will be used during analysis. The analyst should be sure that the Add Health weights are used as sampling weights.

Normalizing the Sampling Weights. Do *NOT* normalize the weights (by dividing the survey weight of each unit used in the analysis by the [unweighted] average of the survey weights of all the analyzed units) unless you are instructed to do so either by the software developer or in documentation supplied with the software. If you normalize the weights, estimates of population totals will be incorrect, even if you use the survey software.

3.2 Steps to Prepare the Data for Analysis

The two main goals of any analysis using data from a complex survey are to produce

1) unbiased estimates of parameters for the entire population as well as subpopulations, and
2) unbiased estimates of variance and standard errors

We have shown that the easiest, quickest, and most reliable way to achieve these two goals when analyzing the Add Health data is to use survey software. It is important then, to select the

appropriate survey software prior to starting data analysis. If you are only interested in the first goal of obtaining unbiased estimates, then you can investigate using your standard statistical analysis package with an appropriate statement to incorporate the sample weights. To obtain unbiased estimates of variance and standard errors, you must account for clustering and correlation of your data.

We next describe necessary steps to prepare the data for analysis. These guidelines have been adapted from "Sampling of Populations: Methods and Applications" (Levy and Lemeshow, 1999).

1. Determine the Wave(s) of data you need for your analysis and construct desired variables.

2. Identify the attributes and elements of the sample design (with replacement Design, strata variable, cluster variable, weight variable) for the data identified in Step 1.

   Design Type: Specify With Replacement as the Design Type

   The information needed to make finite population corrections for analyzing the dataset as a "without replacement design" is not available. However, we can assume that the schools were selected with replacement. The variance estimation technique is derived using large sample theory and will justify our assumption of "with replacement" sampling, even though schools were not placed back on the list before the next school was selected.

   Stratum Variable: REGION

   The Add Health sampling plan did not include a stratification variable. However, a post-stratification adjustment was made to the sample weights so that region of country (variable REGION) could be used as a post-stratification variable. The adjustment involved using the total number of schools in the sampling frame for each region of the country (Northeast, Midwest, South, and West) and for each region adjusting the initial school weights so that the sum of the school weights was equal to the total number of schools in the sampling frame.

   Cluster Variable or Primary Sampling Unit (PSU): PSUSCID

   The variable PSUSCID is the primary sampling unit for the In-School, Wave I, II, III, and IV data. The sampling units in the Add Health Study are middle schools and high schools in the United States. The variable PSUSCID, constructed from the school identifier, is the appropriate variable to use as the cluster or PSU variable.

   Weight Variables

   Determine the type of analysis you intend to do and choose an appropriate weight variable according to the guidelines provided in Chapter 2.

   Note that REGION and PSUSCID variables are located in the same files as weight variables. However, a strata variable is not available for use with the public-use data. Not using a strata

variable only minimally affects the standard errors.

3. Make sure that the variables noted in Step 2 are identified for each sample record.

4. Delete any of the observations that have missing weights from your analysis data set. All of the other design information (strata variable and cluster variable) should be non-missing. Make sure you are analyzing the full sample by checking that the number of observations matches the number given in the tables from Chapter 2. For example, the number of observations in the probability sample from Wave I should be 18,924, and from Wave II should be 13,570.

5. Identify any subpopulation you are interested in analyzing and create an appropriate indicator variable to specify the subpopulation. See Chapter 3, Example 3 for details about using the subpopulation option.

3.3 Variables for Correcting for Design Effects in the Public-Use Dataset

The names for the public-use weight variables differ slightly from the restricted-use names referenced above. In addition to providing the public-use variable names, Table 3.3 includes summary statistics for the public-use weights. Note, a strata variable is not available for the public-use sample but not accounting for the strata with these data only minimally affects the standard errors.

Table 3.3. Public-Use Weight Variables

	Design Type = With Replacement Unit = Adolescent			
	Wave I N=6504	Wave II N = 4834*	Wave III N = 4882*	Wave IV N=5114*
Strata Variable	--------- #	--------- #	--------- #	--------- #
Cluster Variable	CLUSTER2[+]	CLUSTER2[+]	CLUSTER2[+]	CLUSTER2[+]
Weight Variable	GSWGT1	GSWGT2	GSWGT3_2**	GSWGT4_2**
# With Weights	6504	4834	4882	5114
# Missing Weights	0	0	0	0
Mean of Weights	3422.6630	3892.7001	4535.91	4304.66
Sum of Weights	22261000.000	18817312.465	22144327.000	22014038.00
Minimum Weight Value	256.0588	282.4469	295.5669	265.3710
Maximum Weight Value	1835.4864	21107.1003	27327.081	2309.52

Notes.
* These numbers are based on individual datasets, not combined datasets.
# A strata variable is not available; not using a strata variable only minimally affects the standard errors.
[+] The Sociometrics variable name is MEX50197.
** The Wave III and IV files have several weight variables. See chart in codebook to select correct weight to use.

Chapter 4. Software for Analyzing Data from a Sample Survey

There are many software packages available for estimating population-average (marginal or single-level) models from complex survey data. These packages accommodate many different sample designs allowing analysts to adjust for stratification and clustering of observations. Analysts can also specify sampling weights for use during estimation rather than adding covariates to the model to reflect the sampling process. Special features, such as analyzing subpopulations correctly, are available. Recently, software for estimating structural estimation models (SEM) and multilevel models (MLM) have also incorporated many of these same capabilities.

This chapter illustrates the use of several different software packages, primarily SAS and Stata, for estimating population-average models using the Add Health data. We will also provide examples of using several packages to estimate multilevel models, including Mplus, Stata, Lisrel, and MLWin. Illustrative examples are limited to those software packages available at the Carolina Population Center. In Appendix B, we provide SUDAAN syntax for various types of analyses but are unable to provide example results as SUDAAN is unavailable at CPC. Our intent is not to recommend a particular software package, but rather to provide information to our user community. Results from these examples are for the purpose of illustrating the use of the software and may not be representative of actual findings. *These results should not be quoted.*

If you are interested in doing multiple imputation for missing data, you might consider the MI procedure in Stata or IVEware, developed by the Survey Methodology Program at the University of Michigan (http://www.isr.umich.edu/src/smp/ive/). MI in Stata uses the linearization procedure via Taylor series approach, which is sufficient and can account for complex survey features at the estimation level.

IVEware uses variance estimation through the Jackknife approach, which may be necessary in some complex designs, and will produce better variance estimates. IVEware was created with complex survey design in mind. Therefore, it is good software for use with complex survey data. This software can be used to analyze non-normal variables (i.e., proportions, counts, etc.) and can run standard SAS procedures such as PHREG, logistic, and adjust for survey design.

Using STATA for Your Analysis

Stata is an integrated package that offers data management capabilities, and both traditional model-based and design-based analysis capabilities. There is a rich trove of design-based analytical techniques available in Stata. More information is available with the command "help svy." Help with survey commands in Stata is available at http://www.ats.ucla.edu/stat/stata/topics/Survey.htm. Note that some models are not covered by survey methods in Stata and you should refer to the Stata manual for further information.

When employing Stata for design-based analysis, use the command "svyset" to declare survey design features and inform Stata of the design variables you want to include. With the Add Health data, use cluster (primary sampling unit) variable (psuscid), strata variable (region), and weight

variable to specify the survey design characteristics. Stata defaults to a "with replacement" design type, so this information does not need to be specified. The program syntax looks like this:

```
svyset psuscid [pw = wt_var], strata(region)
```

You will need to replace variable "*wt_var*" with weight variables provided in Chapter 2 for single-level models. The choice of the weight variable depends on the type of analysis planned. Several examples of Stata syntax for different types of analysis are provided in the next section.

Example 1. Example for Descriptive Statistics

This example illustrates the use of commands from Stata and SAS to run descriptive statistics. Results from each package are summarized in Table 4.1 and the commands used to estimate the models are listed in Table 4.2.

Research Question: What is the mean number of hours of TV watched during a week for adolescents (data from Wave I in-Home Questionnaire)?

Example 2. Regression Example for Population-Average Models

This example illustrates the use of commands from Stata and SAS that can be used to perform a multiple regression analysis. Results from each package are summarized in Table 4.3. and the commands used to estimate the models are listed in Table 4.4.

Research Question: Is performance on the Add Health Vocabulary Test (PVT_PT1C) influenced by an adolescent's age (AGE_W1), sex (BOY), or time spent watching TV (HR_WATCH)?

Predictive Model:

$$PVT\_PCT1C=\beta_0 + \beta_1 AGE\_W1 + \beta_2 BOY + \beta_3 HR\_WATCH + \text{error term}$$

Where:

β_0 = Intercept
β_1 = Change in Test score for one year increment in age
β_2 =Difference in Test Score between males and females
β_3 =Change in Test Score for each hour spend watching TV

The results are summarized in Table 4.3. Note the results from these packages are nearly identical. Only the standard error for β_0 differs in SAS, but the difference is negligible. The syntax of the program statements for SAS and Stata are given in Table 4.4.

Table 4.1. Parameter Estimates and Standard Errors to Predict the Average Number of Hours TV Watched per Week for Adolescents.

Variable	SAS 9.2.3 Estimate (Std Err)	Stata 12.1 Estimate (Std Err)
hr_tv	15.57 (.36)	15.57 (.36)

Table 4.2. Program Syntax for Descriptive Statistics

Notes: Each program specifies the stratification variable (*region)*, the sampling weight variable (*gswgt1)*, and the cluster (primary sampling unit) variable (*psuscid).* Stata and SAS default to a With Replacement sample.

SAS 9.2.3 syntax:

```
proc surveymeans data=ahw1;
var hr_tv;
cluster psuscid;
strata region;
weight gswgt1;
run;
```

STATA 12.1 syntax:

```
use ahw1.dta, clear
svyset psuscid [pweight=gswgt1], strata(region)
svy: mean hr_tv
```

Table 4.3. Parameter Estimates and Standard Errors to Predict the Percentile Score on the Add Health PVT Test

Parameter	SAS 9.2.3 Estimate (Std Err)	Stata12.1 Estimate (Std Err)
β_0 (INTERCEPT)	69.946 (7.855)	69.946 (7.854)
β_1 (AGE_W1)	-1.085 (0.489)	-1.085 (0.489)
β_2 (BOY)	3.395 (0.673)	3.395 (0.673)
β_3 (HR_WATCH)	-0.150 (0.020)	-0.150 (0.020)

Table 4.4. Program Syntax for Regression Example

Notes: Each program specifies the stratification variable (*region)*, the sampling weight variable (*gswgt1)*, and the primary sampling unit variable (*psuscid).* Stata and SAS default to a "With Replacement" sample. The variable BOY is coded as 0=girl, 1=boy for Stata and SAS.

STATA 12.1 syntax:

```
use ah2006.dta, clear
svyset psuscid [pweight=gswgt1], strata(region)
svy: regress pvtpct1c agew1 boy hr_watch
```

SAS 9.1 syntax:

```
proc surveyreg data=from_w1;
cluster psuscid;
strata region;
weight gswgt1;
model pvtpct1c=agew1 boy hr_watch;
run;
```

Example 3. Subpopulation Analysis

When using survey data, it is common that researchers want to analyze only a certain group of respondents, such as women, those over age 21, or Mexican Americans who reported a history of drug or alcohol use. SUDAAN, Stata, and SPSS all provide special statements or options for analyzing subpopulations using data collected with a complex sampling plan. It is extremely important to use the subpopulation option(s) when analyzing survey data with a sub-sample. If the data set is a subset of the entire Add Health data, i.e., observations not included in the sub-sample/subpopulation are deleted from the data set, the standard errors of the estimates will be wrong. This is because the software needs to be able to identify all PSUs to correctly compute a variance estimate. For example, if a stratum (from the REGION stratification variable) has 132 PSUs and 10 are lost because of restricting the sample to a subset, then the analysis software used to correct for design effects will use an incorrect formula to compute contributions to the variance. When the subpopulation option is used, only the cases defined by the subpopulation are included in the calculation of the estimate, but all cases are included in the calculation of the standard errors (see Cochran, 1977; Rao, 2003).

The magnitude of the difference in the two variance estimates from analyzing the full dataset with the subpopulation option (SUBPOPN, SUBPOP) and the subset of the data is hard to predict. If just a few PSUs are missing in each level of the stratification variable (REGION), then your results will likely be approximately the same. Defining subpopulations by aggregates of the stratification variable in general should not require the subpopulation options be used.

For example, if you wish to analyze all adolescents from REGION=1 level of the stratification variable, you will not need to use the subpopulation option. However, we recommend that you always use the subpopulation options to specify your population of interest. Otherwise, you will have to carefully examine the data to make sure that all PSUs are represented in each level of the stratification variable.

It will often be the case that some of the respondents will not have answered all of the questions included in your analysis. This means that the parameters will not be estimated from the full sample, but rather, from a subset of the data. We recommend that you define the sub-sample of respondents with complete data (no missing on any of the variables) as your subpopulation. This will be particularly useful when you want to compare results from models that contain different subsets of covariates, as you will want the results from all models to be based on the same observations.

Stata example

```
svyset psuscid [pweight=wgt], strata(region)
svy, subpop(nmis): mean v1
```

ID	V1	V2	V3	V4	V5	V6	nmis
1	0	1	2	1	2	0	1
2	1	2	1	0	3	1	1
3	1	3	3	0	1	1	1
4	0	3	4	1	1	0	1
5	.	2	3	.	2	.	0
6	1	.	4	1	.	1	0
7	0	1	.	.	2	0	0
8	0	3	2	0	2	0	1
9	1	1	1	1	3	1	1
10	1	2	4	0	1	0	1

Another scenario is when you use data from multiple panels/waves. For example, you might want to combine data from the Wave I In-School survey (N=83,135), Wave I In-Home survey (N=18,294), and Wave II In-Home survey (N=13,570). After combining the data, the sub-sample size that has data and weights available in all three of these panels would be 10,285. In this case, you need to use subpopulation option to identify a sub-sample of N=10,285.

Before you do the analysis, you should prepare a subpopulation variable. For example, your interest may be in studying a subgroup of Mexican Americans who reported a history of drug or alcohol use. In this case, you would need to create a dummy variable specifying those respondents who belong to this group as 1, and those who do not belong to this group as 0. You would then include this variable in the subpopulation option in your analysis.

In Stata, you could do this for the following sample data:

```
svyset psuscid [pweight=wgt], strata(region)
svy, subpop(mxsub): mean weight
```

ID	race	drug_use	Alcohol_use	weight	mxsub
1	White	No	Yes	120	0
2	Black	Yes	No	140	0
3	Asian	No	Yes	100	0
4	Mexican	Yes	No	135	1
5	Mexican	No	Yes	121	1
6	Asian	Yes	No	115	0
7	Mexican	No	No	140	0
8	White	Yes	No	108	0
9	White	No	Yes	160	0
10	Black	No	Yes	143	0

Note that the subpopulation option is different from the "if" statement. If you use the "if" statement to subset your sample because you are interested in studying a subsample of females, and use "mean weight if bio_sex==2" in Stata, the results will be biased.

Stata has a subpopulation option available. Details about how to use this option in Stata to calculate a mean for this type of subpopulation can be found at: http://www.ats.ucla.edu/stat/stata/faq/svy_stata_subpop.htm. SAS allows users to specify subpopulations with the DOMAIN statement in PROC SURVEYMEANS.

Example 3.1 Example for Descriptive Statistics

Research Question: What is the mean number of hours of TV watched during a week for female adolescents (data from Wave I in-home questionnaire)?

In Table 4.5, we present the results of using SAS and Stata to analyze subpopulations, and in Table 4.6, we show the corresponding syntax.

Table 4.5. Results from using Different Methods of Analyzing Subpopulations

		INCORRECT Deleting cases that are not in subpopulation to subset data	CORRECT Subpopulation option in software	CORRECT DOMAIN statement to specify subpopulation
		Stata 12.1 Estimate (Std Err)	Stata 12.1 Estimate (Std Err)	SAS 9.2.3 Estimate (Std Err)
N of Strata		4	4	4
N of PSUs		131	132	132
N of observations		9582	18870	---
Subpop. No. obs		---	9582	9582
Subpop. size		---	10843943	---
Population size		10843943	---	---
Design DF		127	128	---
Variable	hr_tv	14.55 (.41)	14.55 (.41)	14.55 (.41)

Table 4.6. Syntax for Subpopulation Analysis

Notes: Each program specifies the stratification variable (*region)*, the sampling weight variable (*gswgt1)*, and the primary sampling unit variable (*psuscid).* Stata and SAS default to a With Replacement sample. The variable FEMALE is coded as 1= female 0=male, to specify the female subpopulation.

STATA 12.1 INCORRECT way of subsetting data

Deleting cases that are not in subpopulation to subset data

```
svyset psuscid [pweight=gswgt1], strata(region)
svy: mean tv_hr
```

STATA 12.1 CORRECT way of using SUBPOP option

```
svyset psuscid [pweight=gswgt1], strata(region)
svy, subpop(female): mean tv_hr
```

Alternatively using "over" option for two groups in STATA 12.1: males (0) & females (1)

```
svyset psuscid [pweight=gswgt1], strata(region)
svy: mean tv_hr, over(female)
```

SAS 9.2.3 syntax for using DOMAIN statement to specify subpopulation

```
proc surveymeans data=ahw1;
title3 'Correct subpopulation analysis - set weights to near-zero';
var hr_tv;
cluster psuscid;
strata region;
weight gswgt1;
domain female;
run;
```

Example 3.2 Example for Regression

No other SAS SURVEY procedures allow users to analyze subpopulations. However, the SAS SURVEY software can be tricked into computing the correct variance and standard errors when analyzing subpopulations.

In this section, we illustrate how to implement these tricks by making some slight manipulations of the variables used in the analysis. The example focuses on the research question from the previous section to examine the effect of watching TV on PVT score for adolescents attending rural schools. The model specification is the same as before, however the meaning of the parameter estimates is changed to refer to adolescents attending rural schools. Table 4.7 shows results from different methods of subpopulation analysis. An explanation of each method follows.

Subset Data (INCORRECT). The first second column in Table 4.7 labeled INCORRECT shows results from the wrong method of analyzing subpopulations: subset the data so that observations outside the subpopulation are deleted from the data set being analyzed. Note that this gives the correct parameter estimates, but incorrect standard errors.

Subpopulation option in Software (CORRECT). The third column in the table shows the results using the special statements provided by Stata for analyzing subpopulations. The Stata program statements used to compute these results are shown in Table 4.8. If available in your software package, using the subpopulation option is the best choice for analyzing subpopulation from data collected with a complex survey design. This will ensure that all the details needed to compute estimates, standard errors and test statistics are present and correct.

Set Weights outside the subpopulation Close to Zero. To implement this technique, set the value of the sampling weight close to zero for the sample members who *do not belong* to the subpopulation of interest. This method removes the contribution of an observation to a point estimate, but leaves the structure of the design intact so that the sample survey formulas used to compute variances account properly for the variance in sample size due to potential resampling.

Many software packages, like SAS, delete observations that have a zero value for the sampling weight. In other software packages, a zero value for the weights can lead to numerical errors. One way to avoid these problems is to use a very small weight, rather than zero, to replace the weight for members outside the subpopulation, resulting in estimates that are very close to those computed with a zero weight.

The fourth column in Table 4.7 shows the results from SAS SURVEYREG, where we have used a sampling weight that has a value of 0.00001 for observations outside the population of interest. The estimates are essentially identical to the estimates computed with the subpopulation option in Stata.

Multiply by Subpop Indicator Variable. A second method is to multiply both right and left hand sides of the equation by a subpopulation indicator variable and fit a no-intercept model. In our

example, the subpopulation variable is RURAL (0=non-rural school, 1=rural school). The model from Example 2 becomes:

Predictive Model:

$$\text{RURAL* PVT\_PCT1C} = \beta_0 * \text{RURAL} + \beta_1 (\text{RURAL*AGE\_W1}) + \beta_2 (\text{RURAL*BOY}) + \beta_3 (\text{RURAL*HR\_WATCH}) + \text{error term}$$

The last column in Table 4.7 shows that this method produces the same results as the subpopulation options in Stata. SAS code used for these analyses is shown in Table 4.8.

Table 4.7. Results from using Different Methods of Analyzing Subpopulations

Subpopulation Technique	INCORRECT Subset Data	CORRECT Subpopulation option in software	CORRECT Set Weights outside subpopulation to 0.00001	CORRECT Multiply by Subpop Indicator Variable
Parameter	SAS Estimate (Std Err)	Stata 12.1 Estimate (Std Err)	SAS Estimate (Std Err)	SAS Estimate (Std Err)
β_0 (INTERCEPT)	60.291 (17.40)	60.291 (16.150)	60.291 (16.151)	60.291 (16.151)
β_1 (AGE_W1)	-0.466 (1.08)	-0.466 (1.000)	-0.466 (1.000)	-0.466 (1.000)
β_2 (BOY)	3.409 (1.544)	3.409 (1.445)	3.409 (1.445)	3.409 (1.445)
β_3 (HR_WATCH)	-0.163 (0.03)	-0.163 (0.031)	-0.163 (0.031)	-0.163 (0.031)

Table 4.8. Syntax for Subpopulation Analysis

Notes: Each program specifies the stratification variable (*region)*, the sampling weight variable (*gswgt1)*, and the primary sampling unit variable (*psuscid).* Stata and SAS default to a With Replacement sample. The variable rural is coded as 1= rural school 0=non-rural school. The variable boy is coded as 0=female, 1=male for Stata and SAS.

STATA 12.1 with correct subpopulation option

```
svyset psuscid [pweight=gswgt1], strata(region)
svy, subpop(rural): regress pvtpct1c agew1 boy hr_watch
```

SAS syntax for setting weights to near-zero

```
data from_w1;
set example.ah2006;
rural_wt=gswgt1;
if rural=0 then rural_wt=.00001;
run;

proc surveyreg data=from_w1;
title3 'Correct subpopulation analysis - set weights to near-zero';
cluster psuscid;
strata region;
weight rural_wt;
model pvtpct1c=agew1 boy hr_watch;
run;
```

SAS Indicator Variable Method

```
data from_w1;
set example.ah2006;
rural_pvtpct1c=rural*pvtpct1c;
run;
proc surveyreg data=from_w1;
title3 'Correct subpopulation analysis - multiply both sides by
subpopulation indicator variable';
cluster psuscid;
strata region;
weight gswgt1;
model rural_pvtpct1c=rural rural*agew1 rural*boy rural*hr_watch/noint;
run;
```

Example 4. Multilevel Models

Because of the special attributes of the sample design in Add Health, one can use two levels of data for analysis, including both the school-level and individual level data. With the multi-stage sampling procedure, the probability of selection for both schools and individuals is known. Thus Add Health is able to make two levels of weight components available to users (see Table 2.2). The level 1 weight component pertains to individuals (respondents) and the level 2 weight pertains to PSU (schools). Users who want to use both school-level and individual-level data need to use these two levels of weight components to ensure unbiased population parameters. Note that no neighborhood-level component variable is available in Add Health.

Scaling Sampling Weights

It is important to note that the two level sampling weights should be scaled before running a multi-level model in different packages. Scaling methods may differ depending on the package used. There are two different methods of scaling the sampling weights for estimating this model.

PWIGLS METHOD 2

The first option is to use PWIGLS Method 2 to scale the level 1 weight for the MLM analysis (Pfefferman, 1998). PWIGLS Method 2 is recommended when informative sampling methods are used for selecting units at both levels of sampling. The scaled level 1 weight for each unit *i* sampled from PSU *j* is computed by dividing each level 1 weight by the average of all level 1 weight components in cluster j:

$$pw2r\_w1_{i|j} = \frac{w1\_wc_{i|j}}{\left(\frac{\sum_{i}^{n_j} w1\_wc_{i|j}}{n_j} \right)}$$

There are several packages and procedures that use PWIGLS Method 2 scaling, including XTMIXED and GLLAMM in Stata, MLWIN, and LISREL.

XTMIXED in Stata 12.1 has a "pwscale(size)" option that will automatically use PWIGLS Method 2 to perform the scaling. Therefore, you do not need to use PWIGLS program to do the scaling before you run XTMIXED in Stata 12.1. You simply add the option "pwscale(size)" in XTMIXED and the weights will be automatically scaled.

If you use GLLAMM in Stata to run multi-level models, you need to use PWIGLS (a user written program) to scale the two-level sampling weights before you run GLLAMM. MLWIN and LISREL will automatically do this scaling for the user. In MLWIN, the weights are assumed to be independent of random effects. So you do not need to run PWIGLS to scale weights in these two packages.

Table 4.9. Sampling Weight Scaling and Statistical Packages/Procedures

	Use PWIGLS Method 2	Need to use PWIGLS program to do the scaling before running the multi-level model	Use MPML Method A	Need to use MPML_WT program to do the scaling before running the multi-level model
XTMIXED in Stata	**Yes**	No. Instead, use "pwscale(size)" option in XTMIXED	No	NA
GLLAMM in Stata	**Yes**	**Yes**	No	NA
LISREL	**Yes**	No	No	NA
MLWIN	**Yes**	No	No	NA
HLM*	**Yes**	No	No	NA
MPlus	No	NA	**Yes**	**Yes**

Note: Users of the Add Health data can download SAS and/or Stata programs, PWIGLS and/or MPML_WT for scaling the weights. See Appendix A.

*See Appendix C for scaling in HLM.

MPML METHOD A

A second scaling method is MPML Method A. MPLUS uses weights at both levels of sampling to construct one scaled sampling weight for the two-level analysis. Sampling weights for use with MPLUS two-level models are constructed using MPML Method A. Method A weight construction involves dividing the product of the level 1 and level 2 weight components by the average of the level 1 weight components for units sampled from cluster j:

$$mp\_wt\_w1_{i,j} = \frac{w1\_wc_{i|j} * schwt1_j}{\left(\frac{\sum_i^{n_j} w1\_wc_{i|j}}{n_j} \right)}$$

This computation provides the product of the PWIGLS scaled level 1 weight and the level 2 weight. The analyst must employ the user-written program, MPML_WT, to create the weight for

MPLUS. Table 4.9 shows a summary of how users can scale weights based on the statistical package or procedure used to run the multi-level model.

Example

Data used in this example illustrating the multilevel software packages comes from the School Administrator Survey and the Wave I In-home survey. This example will estimate body mass index of the students in a school from the hours spent watching TV or using computers and the availability of a school recreation center. Information on the availability of an on-site school recreation center (variable RC_S) was provided by each school. Each adolescent answered questions that were used to compute percentile body mass index (BMIPCT) and hours watching TV or playing video or computer games during the past week (HR_WATCH). Our example will fit an MLM with a level for the school and a level for the adolescent. The algebraic formulas describing the model and assumptions follow:

Student-level model (Within or Level 1):

$$(BMIPCT)_{ij} = \{\beta_{0j} + \beta_{1j}(HR\_WATCHij)\} + e_{ij}$$

where:

$$E(e_{ij}) = 0 \quad \text{and} \quad Var(e_{ij}) = \sigma^2$$

School-level Model (Between or Level 2):

$$\beta_{0j} = \gamma_{00} + \gamma_{01}(RC\_S)_j + \delta_{0j}$$

$$\beta_{1j} = \gamma_{10} + \gamma_{11}(RC\_S)_j + \delta_{1j}$$

where:

$$E(\delta_{0j}) = E(\delta_{1j}) = 0, \quad Var(\delta_{0j}) = \sigma^2_{\delta 0}, \quad Var(\delta_{1j}) = \sigma^2_{\delta 1}, \quad Cov(\delta_{0j}, \delta_{1j}) = \sigma_{\delta 01}$$

In this example, we will adjust for the sample design by using the sampling weights to adjust for unequal probability of selection.

The results of the estimation using each package are given in Table 4.10. Lisrel gives estimates that differ from the other packages. We have been notified by the Lisrel developers that there is a problem with the implementation of the multilevel weighting in Lisrel version 8.8 and earlier. Users are advised to use a later version of this software. The program syntax used to compute the results in table 4.10 is given in table 4.11. A similar dataset was created to test the procedure of xtmixed in Stata 12 and run the multi-level model in Mplus, in order to compare the results of the two programs. See Table 4.13 for program syntax used to compute the results presented in Table 4.12.

Table 4.10. Results from Estimation of 2-Level Model Estimated with Sampling Weights

Parameter in 2-Level Model	MPLUS 4.0 Estimate (S.E)	LISREL 8.8 Estimate (S.E.)	MLWIN 2.02 Estimate (S.E.)	GLLAMM Estimate (S.E.)
Weighting method used	MPML Method A	PWIGLS Method 2	PWIGLS Method 2	PWIGLS Method 2
Fixed Effects				
γ_{00} (Intercept for β_{0j})	60.22 (1.09)	59.26 (0.83)	60.28 (1.17)	60.22 (1.10)
γ_{01} (Slope for β_{0j})	-5.48 (1.49)	-3.01 (1.13)	-5.62 (1.65)	-5.48 (1.50)
γ_{10} (Intercept for β_{1j})	0.032 (0.022)	0.043 (0.022)	0.030 (0.023)	0.032 (0.022)
γ_{11} (Slope for β_{1j})	0.13 (0.031)	0.11 (0.028)	0.130 (0.032)	0.13 (0.031)
Random Effects				
$\sigma^2_{\delta 0}$ (Var (δ_{0j}))	19.13 (6.94)	9.16 (1.74)	20.18 (6.04)	19.32 (6.97)
$\sigma^2_{\delta 1}$ (Var (δ_{1j}))	0.003 (0.002)	0.001 (0.001)	0.003 (0.001)	0.003 (0.002)
σ_{12} (Cov (δ_{0j},δ_{1j}))	-0.081 (0.097)	-0.063 (0.034)	-0.091 (0.071)	-0.079 (0.097)
σ^2 (Var (e_{ij}))	788.79 (16.96)	798.15 (76.05)	786.37 (86.62)	788.81 (17.02)

Table 4.11. Program Syntax for Multilevel Analysis

MULTILEVEL ANALYSIS PROGRAM STATEMENTS

```
MPLUS 4.0

*** First, use MPML_WT program to scale the weights (see Appendix A):

DATA:   FILE IS "m:\mp2lev.dat";
        TYPE IS Individual;
VARIABLE:  NAMES ARE aid mp_wt_w1 region psuscid bmipct bmi_qtl bmi_q
        bmi_q4 hr_watch rc_s watch_rc;
        MISSING ARE .;
        USEVARIABLES ARE mp_wt_w1 psuscid bmipct hr_watch rc_s;
        WITHIN = hr_watch;
        BETWEEN = rc_s;
        CLUSTER = psuscid;
        WEIGHT = mp_wt_w1;

ANALYSIS:   TYPE = TWOLEVEL RANDOM;
MODEL:     %WITHIN%
        slope | bmipct ON hr_watch;
        %BETWEEN%
        bmipct slope ON rc_s;
        bmipct WITH slope;
```

GLLAMM (in Stata 9)

*** First, use PWIGLS program to scale the weights (see Appendix A).

*** Note, use original school-level weight component variable for school-level weight; and use *rescaled* individual-level weight variable for individual-level weight.

generate mlwt2=schwt1

generate mlwt1=pw2r_w1

generate one=1

eq sch_int: one

eq sch_slop: hr_watch

gllamm bmipct rc_s hr_watch watch_rc , **i**(sch_id) **nrf(**2**)** ///

eqs(sch_int sch_slop**) pweight(**mlwt**) trace adapt iter(20) nip(12)**

LISREL

***** Do not need to use PWIGLS program to scale weights. It automatically scales the weights.**

OPTIONS OLS=YES CONVERGE=0.001000 MAXITER=10 COVBW=YES OUTPUT=STANDARD ;

TITLE=test;

MISSING_DAT =-9999.000000 ;

MISSING_DEP =-9999.000000 ;

SY='M:\ls2lev4.psf';

ID2=psuscid;

WEIGHT2=schwt_1;

WEIGHT1=w1_wc;

RESPONSE=bmipct;

FIXED=intcept hr_watch rc_s watch_rc;

RANDOM1=intcept;

RANDOM2=intcept watch_rc;

MLWIN (see graphical interface display that follows. Note that the sampling weights are specified with the Weights window accessed from the Model menu. Select "Use standardized weights" for the weighting mode.

***** Do not need to use PWIGLS program to scale weights. It automatically scales the weights.**

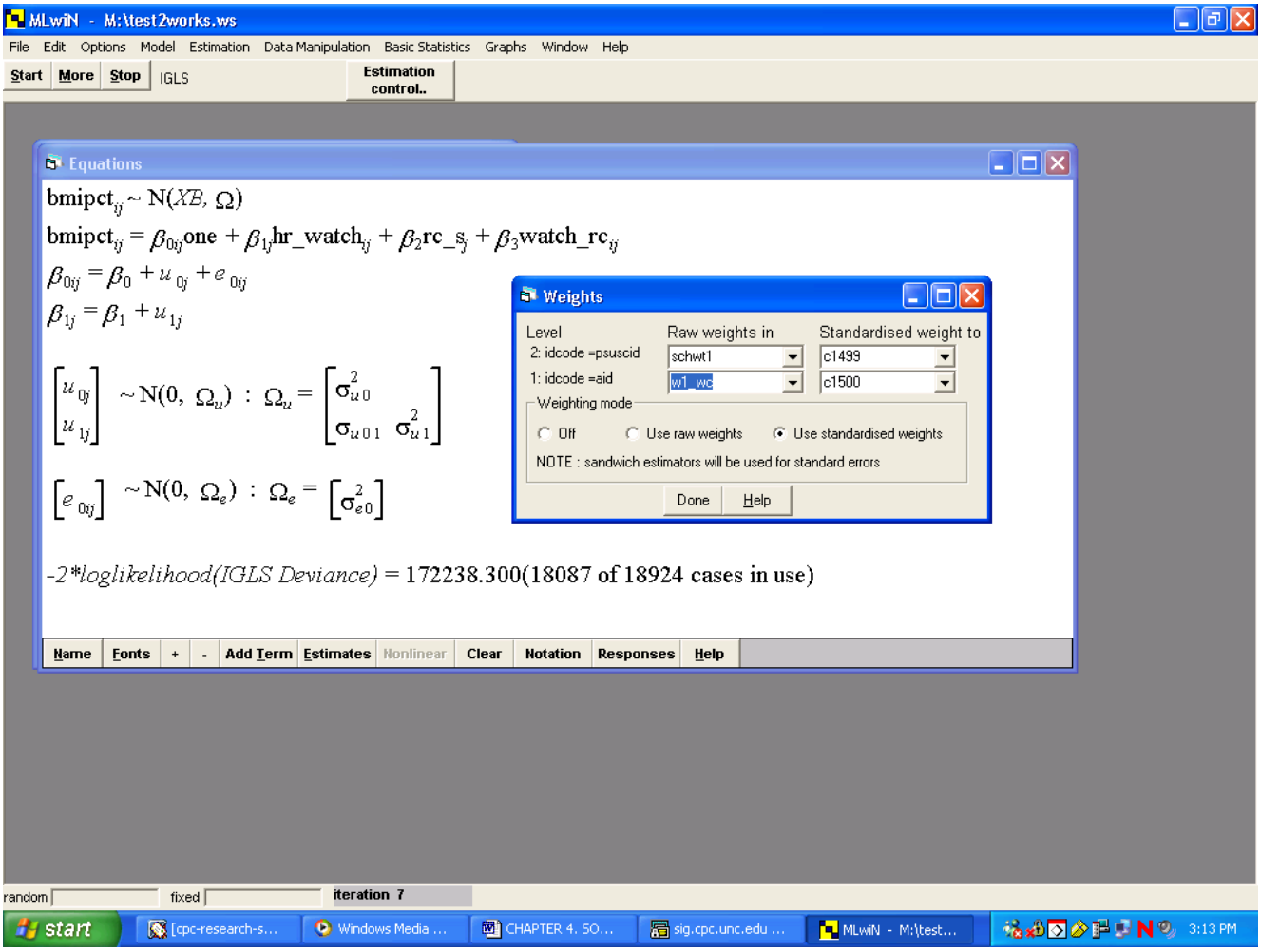

MLwiN - M:\test2works.ws
File Edit Options Model Estimation Data Manipulation Basic Statistics Graphs Window Help
Start More Stop IGLS
Estimation control..
Equations
bmipct_ij ~ N(XB, Ω)
bmipct_ij = β_0ij one + β_1j hr_watch_ij + β_2 rc_s_j + β_3 watch_rc_ij
β_0ij = β_0 + u_0j + e_0ij
β_1j = β_1 + u_1j
[u_0j; u_1j] ~ N(0, Ω_u) : Ω_u = [σ²_u0 ; σ_u01 σ²_u1]
[e_0ij] ~ N(0, Ω_e) : Ω_e = [σ²_e0]
-2*loglikelihood(IGLS Deviance) = 172238.300(18087 of 18924 cases in use)
Weights
Level
Raw weights in
Standardised weight to
2: idcode =psuscid
schwt1
c1499
1: idcode =aid
w1_wc
c1500
Weighting mode
Off
Use raw weights
Use standardised weights
NOTE : sandwich estimators will be used for standard errors
Done
Help
Name Fonts + - Add Term Estimates Nonlinear Clear Notation Responses Help
random
fixed
iteration 7
start
[cpc-research-s...
Windows Media ...
CHAPTER 4. SO...
sig.cpc.unc.edu ...
MLwiN - M:\test...
3:13 PM

Table 4.12. Results from Estimation of 2-Level Model Estimated with Sampling Weights.

Parameter in 2-Level Model	MPLUS 4.0 Estimate (S.E)	XTMIXED Estimate (S.E.)
Weighting method used	MPML Method A	PWIGLS Method 2
Fixed Effects		
γ_{00} (Intercept for β_{0j})	0.458 (0.009)	0.450 (0.012)
γ_{01} (Slope for β_{0j})	-0.025 (0.015)	-0.049 (0.030)
γ_{10} (Intercept for β_{1j})	0.000 (0.000)	0.000 (0.000)
γ_{11} (Slope for β_{1j})	0.001 (0.000)	0.001 (0.000)
Random Effects		
$\sigma^2_{\delta 0}$ (Var (δ_{0j}))	0.005 (0.001)	0.005 (0.001)
$\sigma^2_{\delta 1}$ (Var (δ_{1j}))	0.000 (0.000)	0.000 (0.000)
σ_{12} (Cov (δ_{0j},δ_{1j}))	0.000 (0.000)	- 0.000 (0.000)
σ^2 (Var (e_{ij}))	0.074 (0.001)	0.077 (0.002)

Table 4.13. Program Syntax for Multilevel Analysis

MULTILEVEL ANALYSIS PROGRAM STATEMENTS

MPLUS 4.0

*** First, use MPML_WT program to scale the weights (see Appendix A):

```
DATA:   FILE IS "d:\xtmixed_test.dat";
        TYPE IS Individual;
VARIABLE:  NAMES ARE aid psuscid region w1bmirk w1hr_tv w1rc w1tv_rc
                     mp_wt_w1;
        MISSING ARE ALL (-9999);
        USEVARIABLES ARE mp_wt_w1 psuscid w1bmirk w1hr_tv w1rc;
        WITHIN = w1hr_tv;
        BETWEEN = w1rc;
        CLUSTER = psuscid;
        WEIGHT = mp_wt_w1;

ANALYSIS:   TYPE = TWOLEVEL RANDOM;
MODEL:      %WITHIN%
        slope | w1bmirk ON w1hr_tv;
        %BETWEEN%
        bmipct slope ON w1rc;
        w1bmirk WITH slope;
```

XTMIXED (in Stata 12.1)

***** option "pwscale(size)" automatically uses PWIGLS Method 2 to scale the two-level weights.**

```
xtmixed w1bmirk w1rc w1hr_tv w1tv_rc [pw=w1_wc] ///
        || psuscid: w1hr_tv, pweight(schwt1) pwscale(size) nolog  var cov(unst)
```

Appendix A. Scaling Weights for Multilevel Analysis

User-written Stata and SAS programs for scaling sampling weights to estimate two-level models that can be used with several popular multilevel software packages can be downloaded from our website:

http://www.cpc.unc.edu/research/tools/data_analysis/ml_sampling_weights

Also available from the CPC website (http://www.cpc.unc.edu/research/tools/data_analysis) is documentation that provides (1) information on using these programs to create the two-level weights, (2) information about several popular multilevel software packages that allow these sampling weights to be used in estimation, and (3) instructs the analyst in downloading and running these programs.

Users of gllamm and Mplus 4.1 and earlier will need to scale the weights as described above, in Example 4 on multilevel models. Users of these programs can scale the weights by writing their own program or by using the SAS and Stata programs provided on the CPC website. The statements using these programs are included in the following tables.

Table A1. Example code used to construct weights for gllamm used in Example 3

PWIGLS METHOD OF WEIGHT CONSTRUCTION FOR EXAMPLE 3

SAS PWIGLS Macro

```
%include '/bigtemp/sas_macros/pwigls.sas';
%pwigls(input_set=testdat,
        psu_id=psuscid,
        psu_wt=schwt1,
        fsu_id=aid,
        fsu_wt=w1_wc,
        output_set=pwigl_wt,
        psu_m1wt = pw1s_w1adj,
        fsu_m1wt = pw1r_w1,
        psu_m2wt = pw2s_w1adj,
        fsu_m2wt = pw2r_w1,
        replace=replace);
run;
```

STATA PWIGLS Command

```
use testdat, clear
pwigls, psu_id(psuscid) fsu_id(aid) psu_wt(schwt1) fsu_wt(w1_wc) psu_m1wt(m1adj)
fsu_m1wt(pw1r_w1) psu_m2wt(m2adj) fsu_m2wt(pw2r_w1)
```

Detailed instructions on running this software and definitions of variables can be found in the previously mentioned documentation, available on the CPC website. The variables psuscid (identifying the school), the level 2 weight component (schwt1), the respondent identifier (aid), and the level 1 weight component (w1_wc), should be in the input data set (testdat). The PWIGLS program will return weights scaled by both methods. Only the PWIGLS method 2 weight scaled weight is needed for analysis. In this example, the weight is called pw2r_w1 and is the scaled level 1 weight required by gllamm.

Users of MPLUS 4.1 may use the PWIGLS macro and multiply the level 2 weight and PWIGLS scaled level 1 weight together to produce the required combined weight. For this example, the MPLUS combined weight is calculated as:

```
mp_wt_w1 = pw2r_w1*schwt1
```

Alternatively, users can download the MPML_WT programs that will scale the weights according to the instructions given in Example 4, above.

Table A2. Example Code used to Construct Composite Weight for MPLUS used in Example 4.

WEIGHT CONSTRUCTION FOR MPLUS

SAS MACRO FOR MPLUS COMPOSITE WEIGHT
`%include '/bigtemp/sas_macros/mpml_wt.sas';`
`%mpml_wt(input_set=testdat,`
`psu_id = psuscid,`
`fsu_id = aid,`
`psu_wt = schwt1,`
`fsu_wt= w1_wc,`
`output_set = mpml_dat,`
`mpml_wta = mp_wt_w1,`
`replace=replace);`

STATA COMMAND FOR MPLUS COMPOSITE WEIGHT `mpml_wt, psu_id(psuscid) fsu_id(aid) psu_wt(schwt1) fsu_wt(w1_wc) mpml_wta(mp_wt_w1)`

The variables psuscid (identifying the school), the level 2 weight component (schwt1), the respondent identifier (aid), and the level 1 weight component (w1_wc) should be in the input data set (testdat). The option mpml_wta will generate the weight variable "mp_wt_w1" for use in estimating 2-level models in Mplus.

Appendix B. SUDAAN Syntax for Different Types of Analysis

Using SUDAAN for Your Analysis

SUDAAN template takes the form:

```
PROC whatever data="AH_data" FILETYPE=SAS DESIGN=WR;
NEST REGION PSUSCID;
WEIGHT wt_var;
SUBPOPN mydata=1;
Add other modeling statements, printing options here;
```

The first statement specifies the appropriate SUDAAN procedure for your analysis, the name (*AH_data*) and type (*SAS*) of the data file, and indicates the appropriate design, "with replacement" (WR). You will need to replace *whatever* with the procedure name. The second statement (NEST command) specifies the strata variable (REGION) and primary sampling unit or cluster variable (PSUSCID). Unless otherwise specified, SUDAAN assumes the first variable in this statement is the stratification variable and the second is the primary sampling unit. The fourth statement is used to specify the population of interest for your analysis. The variable *mydata* is an indicator variable, with a value of 1 for all observations that need to be included in the parameter estimates and 0 for observations you want omitted.

The variable boy_r is coded as 1=male, 2=female for SUDAAN. SUDAAN requires the variable identifying the PSU to be numeric, so psuscidn is a numeric version of the Add Health character variable PSUSCID.

Program Syntax for Descriptive Analysis:

```
proc descript data="ALLKIDS" filetype=SAS design=WR;
nest region psuscidn;
weight gswgt2;
var hr_tv ;
setenv pagesize=40 linesize=60;
title "USE ALLKIDS for Descriptive Analysis";
run;
```

Program Syntax for Regression Example:

```
proc regress data=”from_w1” filetype=SAS design=WR
semethod=binder;
nest region psuscidn;
weight gswgt1;
class boy_r;
model pvtpct1c=agew1 boy_r hr_watch;
run;
```

Program Syntax for Descriptive Statistics and Subpopulation Analysis:

```
proc descript data=”ALLKIDS” filetype=SAS design=WR;
nest region psuscid;
weight gswgt2;
subpopn rural=1;
var hr_tv ;
setenv pagesize=40 linesize=60;
title “USE ALLKIDS with SUBPOPN statement”;
```

Program Syntax for Regression and Subpopulation Analysis:

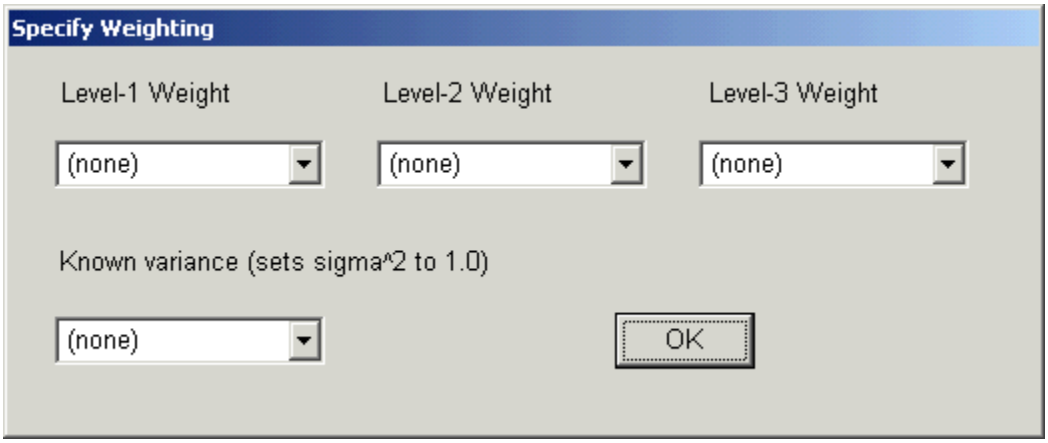
```
proc regress data=”from_w1” filetype=SAS design=WR
semethod=binder;
title3 'Correct subpopulation analysis in SUDAAN';
nest region psuscidn;
subpopn rural=1;
weight gswgt1;
class boy_r;
model pvtpct1c=agew1 boy_r hr_watch;
print /betafmt=f10.6 sebetafmt=f10.6;
run;
```

Appendix C. Incorporating Two-Level Weight Components in HLM

The following information is based on:

http://www.ssicentral.com/hlm/example6-2.html

To analyze two-level data in HLM v6, weights are selected at the time of analysis, rather than when the MDM file is constructed. To select weights for an HLM2 analysis (two-level linear and nonlinear [HGLM] models), select the **Estimation Settings** option from the **Other Settings** menu, and use the pull down menus to select the weighting variables at any level.



Enter the level-1 weight component variable (listed in column 3 table 2.2) as the "Level-1 Weight" option and the level-2 weight component variable (listed in column 2 table 2.2) as the "Level-2 Weight" option. HLM will then automatically use PWIGLS Method 2 to perform the scaling.

Appendix D. Svyset Add Health Data with Two-Level Cross-Sectional Modeling in Stata 14

We have discussed how to use and scale weight component variables for two-level models in Chapter 4 Example 4 and Appendix A-C in this paper to account for the unequal probability of sample selection.

In prior versions, Stata was not able to svyset the weight, clustering and stratification variables and use svy prefix for multilevel models. Stata 14 starts to integrate svyset command and svy prefix to account for features of complex survey design.

The goal of this appendix is to show how to **apply svyset command** in a **two-level cross-sectional context** to account for Add Health survey design features, including unequal probability of selection, clustering, and stratification.

If you are using both school-level and individual-level data to estimate a two-level cross-sectional model, you could consider the new survey procedures in Stata 14. If you use ***svyset*** command to set up your two-level weight component variables, you do ***not*** need to scale the two component variables.

Stata provides a brief summary of survey support for multilevel models at:

http://www.stata.com/new-in-stata/multilevel-models-survey-data/

We mimic the example given by Stata to show how you could apply this to the Add Health context.

Variable list for estimating a two-level cross-sectional logistic model with Add Health Wave I data		
Variables of interest	Variable Value & Label	Variable Name
WI Y	1=obese; 0 = not obese	w1obese
WI School-level X	1=school-level recreation center available 0 = not available	w1schrecrctr
WI Individual-level X	Number of hours spent by respondents watching TV	w1hr_tv
WI weight component variables		
School level		schwt1
Individual-level cross-sectional		w1_wc
Cluster variable		psuscid
Stratification Variable		region
Subpopulation variable	1= not missing in any of the variables included in the model 0 = missing in one or more of the variables of interest	nonmiss

Full Sample:

```
svyset psuscid, weight(schwt1) strata(region) || aid, weight(w1_wc)
svy: melogit w1obese w1hr_tv w1schrecrctr || psuscid:
```

Sub-sample of respondents who are not missing in any of the variables included in the svyset command and svy estimation procedure:

```
svyset psuscid, weight(schwt1) strata(region) || aid, weight(w1_wc)
svy, subpop(nonmiss): melogit w1obese w1hr_tv w1schrecrctr || psuscid:
```

Things to note:

1. **Choose a single-level model and single-level weight** if you are only interested in including school-level variables as covariates but not in obtaining variance components estimates (i.e. random effects).

2. **Choose a single-level model and single-level weight** if you are only interested in estimating **population average** even when you are using longitudinal data.

3. Add Health does *not* provide 3-level weights if you estimate a three-level model.

4. Stata provides a comprehensive online documentation of users' manual. Please refer to Stata manual for model specification questions. Or you could always contact tech-support@stata.com for support.

References Cited

Chantala K, Blanchette D, Suchindran CM. *Software to compute sampling weights for multilevel analysis*. Carolina Population Center, University of North Carolina at Chapel Hill, 2011. Available at http://www.cpc.unc.edu/research/tools/data_analysis/ml_sampling_weights/Compute%20Weights%20for%20Multilevel%20Analysis.pdf

Cochran WG. Sampling Techniques, 3rd Edition. Cambridge, MA: Harvard University, 1977.

Harris KM. The Add Health Study: Design and Accomplishments. Carolina Population Center, University of North Carolina at Chapel Hill, 2013. Available at http://www.cpc.unc.edu/projects/addhealth/data/guides/DesignPaperWIIV.pdf

Levy PS, Lemeshow S. Sampling of populations: methods and applications. John Wiley & Sons, 1999.

Rao JNK. *Small Area Estimation*. Wiley Series in Survey Methodology, 2003.

Tourangeau R, Shin H-C. National Longitudinal Study of Adolescent Health Grand Sample Weight. Carolina Population Center, University of North Carolina at Chapel Hill, 1999. http://www.cpc.unc.edu/projects/addhealth/data/guides/weights.pdf

Additional Information

1. Websites

   Add Health: http://www.cpc.unc.edu/projects/addhealth

   Centre for Multilevel Modeling: http://www.bristol.ac.uk/cmm/

   MPLUS: http://www.statmodel.com/

   SUDAAN: http://www.rti.org/sudaan/

   STATA: http://www.stata.com/

   SAS: http://www.sas.com/

2. Information about survey software packages:
   https://www.stattransfer.com/stattransfer/formats.html

3. List Servers

   Add Health: to interact with other data users and analysts: Send email to listserv@unc.edu and in the body of the message type:

   subscribe addhealth2 *firstname lastname*

   Add Health: to receive notifications about data and documentation: Send email to addhealth@unc.edu and in the subject line put *Add Health List Server*.

4. Supplemental Reference Material

   Asparouhov T. *Sample weights in latent variable modeling*. Muthen and Muthen, Mplus Webnotes 72, 2005. Available at http://www.statmodel2.com/download/webnotes/mplusnote72.pdf

   Asparouhov T. *Weighting for unequal probability of selection in multilevel modeling*. Muthen and Muthen, Webnote 8, 2004. Available at http://www.statmodel.com/download/webnotes/MplusNote81.pdf

   Brogan D, Daniels D, Rolka D, Marsteller F, Chattopadhay M. Software for sample survey data: misuse of standard packages. Invited Chapter in Encyclopedia of Biostatistics, P. Armitage and T. Colton (eds.), Vol 5, pp. 4167-4174. John Wiley, New York, 1998

   Chantala K, Tabor J. *National Longitudinal Study of Adolescent Health: Strategies to perform a design-based analysis using the Add Health data*. University of North Carolina at Chapel Hill, 1999.

   Cohen SB. An evaluation of alternative PC-based software packages developed for the analysis of complex survey data. *The American Statistician*, August 1997, Vol. 51, No. 3, pages 285-292.

   Goldstein H. Multilevel Statistical Models, Kendall's Library of Statistics 3. London Institute of Education, 1999. Internet edition, available at http://www.soziologie.uni-halle.de/langer/multilevel/books/goldstein.pdf

   Littell RC, Milliken GA, Stroup WW, Wolfinger RD. SAS System for Mixed Models, Cary,

NC: SAS Institute, 1996.

Muthén L, Muthén B. Mplus User's Guide, Los Angeles, CA, 2000.

SAS Institute Inc. SAS/STAT Software: Changes & Enhancements through Release 6.12. Cary, NC: SAS Institute, 1997.

Shah BV, Barnwell BG, Bieler GS. SUDAAN User's Manual: Release 6.4. Research Triangle Institute: Research Triangle Park, NC, 1995.

Singer J. Using SAS PROC MIXED to fit multilevel models, hierarchical models, and individual growth models. Available at http://www.gse.harvard.edu/~faculty/singer/Papers/Using%20Proc%20Mixed.pdf

Stapleton LM. The incorporation of sample weighs into multilevel structural equation models. *Structural Equation Modeling* 2002; 9(4), 475-502.

Stata Corporation. Stata Reference Manual, Release 6. College Station, TX, 1999.

Williams RL. A note on robust variance estimation for cluster-correlated data. *Biometrics* 2000; 56, 645-646.