# Guidelines for Analyzing Add Health Data

Kim Chantala
Carolina Population Center
University of North Carolina at Chapel Hill

Last Update: October 1, 2006

# Table of Contents

# Chapter 1.  Basic Concepts of the Add Health Design

The sampling plan used to collect the Add Health data has resulted in the Add Health sample differing from the target population of U.S. adolescents in ways that can influence analysis results.  This section describes how the Add Health sample was selected and discusses the 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 for 1994 - 1995 academic year.  A sample consisting of 132 schools was chosen with unequal probability of selection.  First, a list of 26,666 U.S. High Schools was sorted on enrollment size, school type, region, location, and percent white and then divided into groups for sampling.  Eighty high schools were selected systematically from this list with probability proportional to enrollment size.  High schools that did not include 7th or 8th grades supplied names of middle schools that contributed students to the incoming class.   For each of these high schools, a single feeder school was selected with probability proportional to the percentage of the high schools' entering class that came from the feeder school.  A total of 52 feeder (junior high & middle) schools were selected.  Administrators at each school were asked to fill out a special survey that captured attributes of the school.

Add Health has collected multiple panels of data on adolescents recruited from these schools. These panels are as follows:

- In-School Survey (1994):  All students who were in attendance on day of interview were asked to fill out a questionnaire.

- Wave I In-Home Survey (1995): Adolescents were selected with unequal probability of selection from the 1994-1995 enrollment rosters for the schools.  These include the following subpopulations:  the core sample (roughly equal-sized samples selected from most schools), purposively selected schools, over sampled groups of adolescents with specific racial or ethnic backgrounds, over-sampled group of disabled youth, and the genetic supplement.

- 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. 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.

A detailed list of attributes for selecting schools and adolescents appears in Table1.1.

Table1.1. Attributes of the Add Health sampling design influencing an adolescent being selected for recruitment are listed below.

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

All of the attributes listed in Table1.1 used to select both Add Health schools and adolescents, as well as characteristics related to non-response, have been used 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.

## 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 that influenced being a participant in the Add Health Study also influences the outcome of interest. For example, black adolescents whose parents were college graduates are one of the many over-sampled groups. Thus, parental education is a factor that affected participation of black youth in the Add Health study and can influence family income. Unless the analysis 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 can also 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 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. Table1.2 lists the attributes of the Add Health sampling design that should be taken into consideration during analysis.

Table1.2. Add Health sampling design attributes.

| Design Attribute | Usual Impact on Analysis | Variables Provided with Add Health Data 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:<br>• Cross-sectional Weights for Schools.<br>• Cross-sectional Weights for analyzing each Wave of Data.<br>• Cross-sectional Weights for analyzing special sub-samples from Wave III.<br>• Longitudinal Weights for analyzing analyses combining data form multiple Waves.<br>• 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.  The first category includes analyses to provide population estimates for adolescents who were enrolled in secondary school for the 1994-1995 academic year (Table 2.1).  Often these analyses involve fitting a population-average (single-level or marginal) model.

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) | Sample | Target Population |
|---|---|---|---|
| Wave I (1995) | GSWGT1 (N=18,924) | 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) | 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) | Wave I respondents who were interviewed at Wave III. | Grade 7-12 [1] in 1994-1995 |
| Wave III (2001) | GSWGT3 (N=10,828) | Eligible Wave I Respondents interviewed at both Wave II & Wave III. | Grade 7-11 [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

The second category includes analyses fitting a multilevel-model to provide estimates for adolescents who were in secondary 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 the basic building blocks needed for computing the multilevel weights with the methods given in the web document located at: *http://www.cpc.unc.edu/restools/data_analysis/ml_sampling_weights*

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

| Interview (Year collected) | Level 2 Weight Component (N) | Level 1 Weight Component (N) | Sample | Target Population |
|---|---|---|---|---|
| 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 |
| Wave 1 (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 |
| 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 |
| 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 |
| 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 |

The third category includes analyses fitting a population-average model for special subpopulations of the US in 2001 enrolled in secondary 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 binge sample includes participants selected at Wave III to study binge-drinking attitudes among college-age students. The eligibility criteria for 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 to 93 in section 28 were asked to approximately equal numbers of respondents who met the eligibility criteria in each of the following four groups:  Females attending college, Males attending college, Females not attending college, Males not attending college.

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) | BINGEWT (N=1,499) | Wave III Binge Sample: Eligible Wave I Respondents asked questions 50 to 93 in Section 28 at Wave III. | College Age Youth in 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 II respondents interviewed at Wave III. | Grade 7-11 [1] in 1994-1995 |
| | *to be supplied* | HPV MGEN Sample:  special sample selected for testing urine for mycoplasma genitalium and Human Papillomavirus. | *to be supplied* |

[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 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 the student's transcript, or provided incomplete or erroneous 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.

### Choosing a Sampling Weight for Analysis

The sampling weight selected for an analysis depends on the type of analysis needed to investigate a hypothesis and which interview or combination of interviews needed in the analysis. This section gives instruction on selecting the best sampling weight for most analyses.

### Cross-Sectional Analysis

Research questions investigated by cross-sectional analysis tend to investigate association rather than causation. The temporal sequence of events necessary for drawing causal inferences may not be available. Data for both predicting and outcome variables are collected at the same point in time. The outcome can be observed for all subjects. The choice of sampling weight will be the weight created for for everyone in the probability sample for the Wave of data you are using (Table 2.4).

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<br>W1_WC |
| Adolescents in 1996 enrolled in Grade 7-11 during 1994-1995 | Wave II | 13,570 | GSWGT2 | SCHWT1<br>W2_WC |
| Young Adults in 2001 enrolled in Grade 7-12 during 1994-1995 | Wave III | 14,322 | GSWGT3_2 | SCHWT1<br>W3_2_WC |
| Young Adults in 2001 enrolled in Grade 7 & 8 during 1994-1995 – (Analyses involving binge drinking habits) | Wave III | 1,499 | BINGEWT | Not Available |
| of Young adults Romantic Couplers 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 |

## **Longitudinal Analysis**

Longitudinal analysis is used to investigate research questions answered by investigating changes in measurements taken on subjects over time. The outcome can be observed for all subjects. The data being analyzed can be organized in different ways. Two common ways are:

- one record per subject (AID) per time point
- records for a subject 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 subject 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 time point because the subject was not interviewed. The analyst only then needs to consider the effect of item non-response rather than both item and survey non-response.

Longitudinal analysis with the Add Health data will involve using information collected at multiple interviews. In general, the choice of sampling weight for longitudinal analysis will be determined by the data collected at the latest time-point. Table 2.5 shows the appropriate sampling weight to use for most longitudinal analyses that estimate population-average models.

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<br>W2_WC |
| Adolescents enrolled in Grade 7-11 during 1994-1995 interviewed in 1995 & 2001 | Wave I & III | 14,322 | GSWGT3_2 | SCHWT1<br>W3_2_WC |
| Adolescents enrolled in Grade 7-11 during 1994-1995 interviewed in 1996 & 2001 | Wave II & III | 10,828 | GSWGT3 | SCHWT1<br>W3_WC |
| Adolescents enrolled in Grade 7-11 during 1994-1995 interviewed in 1995, 1996 & 2001 | Wave I, II, & III | 10,828 | GSWGT3 | SCHWT1<br>W3_WC |

## Time-to-Event Analysis

Research questions best answered by time-to-event analysis involve the occurrence and timing of events. Data involves individuals observed over time where the outcome is the occurrence of a specific event. The event 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 subjects. Choice of sampling weight will usually be determined by the data collected at the earliest time point.

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<br>W1_WC |
| Adolescents in 1996 enrolled in Grade 7-11 during 1994-1995 | Wave II only | 13,570 | GSWGT2 | SCHWT1<br>W2_WC |
| Young Adults in 2001 enrolled in Grade 7-12 during 1994-1995 | Wave III only | 14,322 | GSWGT3_2 | SCHWT1<br>W3_2_WC |
| *Data available from Multiple interviews:* | | | | |
| Adolescents in 1995 enrolled in Grade7-12 during 1994-1995 | Wave I & II | 18,924 | GSWGT1 | SCHWT1<br>W1_WC |
| Adolescents in 1996 enrolled in Grade 7-11 during 1994-1995 | Wave II & III | 13,570 | GSWGT2 | SCHWT1<br>W2_WC |
| Young Adults in 2001 enrolled in Grade 7-12 during 1994-1995 | Wave I, II, & III | 18,924 | GSWGT1 | SCHWT1<br>W1_WC |

## Summary

There guidelines presented in this chapter on 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.

# Chapter 3.  Avoiding Common Errors

This chapter lists the most common errors made when analyzing the 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.  The easiest way to adjust estimates for clustering and unequal probability of selection is to use software that adjusts for clustering and uses sampling weights when computing point estimates and standard errors.  If the software you are using does not allow you to specify sampling weights then consider including covariates in your analysis that are related to schools and adolescents being selected for participation in the Add Health Survey.  These are listed in Table 1.1.

***Including respondents who are missing sampling weights in analyses when your goal is obtaining 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 of 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 (the 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, suppose you are interested in performing an analysis on Asians only.   The sample of students selected from some schools might not include any Asians.  If the sampling was repeated Asians might be selected from these schools and the schools would remain in the analysis.  This variation in school sample size that would occur in re-sampling must be included in estimating variances and standard errors.  Subsetting the data 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 to be used for 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 subjects who were actually interviewed.  For example, a frequency weight of 3 means that the three subjects were interviewed and all gave identical answers to every question.

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

***Sampling Weights***. These weights are computed as the inverse of the probability of selection that this subject 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 subjects in the population of interest.

Each of these weights enters the computation in a different way and will give different estimates 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 the Add Health weights are used as sampling weights.

***Normalizing the Sampling Weights.*** Do not normalize the weights unless you are instructed to by the developers of the software or documentation supplied with the software. If you normalize the software, estimates of population totals will be incorrect even if you use the survey software.

### 3.2 Preparing Your Data for Analysis

These guidelines have been adapted from "Sampling of Populations: Methods and Applications" by Paul S. Levy and Stanley Lemeshow, 1999, John Wiley and Sons.

1. Determine the Wave(s) of data you need for your analysis and construct desired variables.
2. Identify the attributes & elements of the sample design (With Replacement Design, Strata variable, Cluster variable, Weight variable) for the data identified step 1.
3. Make sure that the above variables identified in step 2 are identified on 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 use for specifying the subpopulation.

# 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 that reflect the sampling process. Special features, such as the analyzing subpopulation 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 using several different software packages for estimating population-average and multilevel models using the Add Health data. Use of a software packages reflects availability at the Carolina Population Center. Our intent is not to recommend a particular software package, but to provide information for our user community. Results from these examples are for illustrating usage of the software and may not be representative of actual findings. These results should not be quoted.

**Example 1. Regression Example for Population-Average Models.**

This example illustrates the use of commands from Stata, SUDAAN, and SAS that can be used to perform a multiple regression analysis. 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:* 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:*

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

*Where:*

$\beta_0$ = Intercept

$\beta_1$ = Change in Test score for one year increment in age

$\beta_2$ =Difference in Test Score between males and females

$\beta_3$ =Change in Test Score for each hour spend watching TV

The results are summarized in Table 4.1. Note the results from these packages are nearly identical. Only the standard error for $\beta_0$ differs in SAS. This difference is negligible. The syntax of the program statements for SAS, Stata and SUDAAN is given in Table 4.2.

Table 4.1 Parameter estimates and standard errors to predict the percentile score on the Add Health PVT test.

| Parameter | SAS 9.1<br>Estimate (Std Err) | Stata 9.2<br>Estimate (Std Err) | SUDAAN<br>Estimate (Std Err) |
|---|---|---|---|
| $\beta_0$ (INTERCEPT) | 69.946 (7.855) | 69.946 (7.854) | 69.946 (7.854) |
| $\beta_1$ (AGE_W1) | -1.085 (0.489) | -1.085 (0.489) | -1.085 (0.489) |
| $\beta_2$ (BOY) | 3.395 (0.673) | 3.395 (0.673) | 3.395 (0.673) |
| $\beta_3$ (HR_WATCH) | -0.150 (0.020) | -0.150 (0.020) | -0.150 (0.020) |

Table 4.2 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 sampled while design=WR is specified in the SUDAAN code. The variable boy is coded as 0=female, 1=boy for Stata and SAS while boy_r is coded as 1=boy, 2=female for SUDAAN.

```
STATA 9.2 syntax:
use ah2006.dta, clear
svyset [pweight=gswgt1], strata(region) psu(psuscid)
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;
```

```
SUDAAN:

proc regress data=from_w1 filetype=SAS design=WR semethod=binder;
nest region psuscidn;
weight gswgt1;
subgroup boy_r;
levels 2;
model pvtpct1c=agew1 boy_r hr_watch;
run;
```

**Example 2. Subpopulation Analysis**

SUDAAN, Stata, and SPSS all provide special statements or options for analyzing subpopulations using data collected with a complex sampling plan. SAS does allow users to specify subpopulations with the DOMAIN statement in PROC SURVEYMEANS. However, none of the 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.

This 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 4.3 shows results from different methods of subpopulation analysis. An explanation of each method follows.

***Subset Data (INCORRECT).*** The first method in the table 4.3 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 standard errors that are incorrect.

***Subpopulation option in Software (CORRECT).*** The next column shows the results using the special statements provided by SUDAAN and Stata for analyzing subpopulations. The Stata and SUDAAN program statements used to compute these results is shown in table 4.4. 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 of computing estimates, standard errors and test statistics are right.

***Set Weights outside the subpopulation to Zero.*** To implement this technique, set the value of the sampling weight 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 use this technique and avoid these problems is to use a very small weight instead of zero to replace the weight for members outside the subpopulation so that the estimates are very close to estimates computed with a zero weight.

The column fourth column in Table 4.3 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 SUDAAN and Stata. Table 4.4 shows the SAS code used for this analysis.

Table 4.3 Results from using different methods of analyzing subpopulations.

| Subpopulation Technique | INCORRECT Subset Data | CORRECT Subpopulation option in software | Set Weights outside subpopulation to Zero | Multiply by Subpop Indicator Variable |
|---|---|---|---|---|
| Parameter | SAS<br>Estimate (Std Err) | Stata 9.2 ,<br>SUDAAN<br>Estimate (Std Err) | SAS<br>Estimate (Std Err) | SAS<br>Estimate (Std Err) |
| $\beta_0$ (INTERCEPT) | 60.291 (17.40) | 60.291 (16.150) | 60.291 (16.151) | 60.291 (16.151) |
| $\beta_1$ (AGE_W1) | -0.466 (1.08) | -0.466 (1.000) | -0.466 (1.000) | -0.466 (1.000) |
| $\beta_2$ (BOY) | 3.409 (1.544) | 3.409 (1.445) | 3.409 (1.445) | 3.409 (1.445) |
| $\beta_3$ (HR_WATCH) | -0.163 (0.03) | -0.163 (0.031) | -0.163 (0.031) | -0.163 (0.031) |

***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 1 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 4.3 shows that this method produces the same results as the subpopulation options in SUDAAN and Stata.

Table 4.4 shows the program statements used to compute the results in Table 4.3.

Table 4.4 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 sampled while design=WR is specified in the SUDAAN code. The variable rural is coded as 1= rural school 0=non-rural school. The variable boy is coded as 0=female, 1=boy for Stata and SAS while boy_r is coded as 1=boy, 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.

**STATA 9.2**

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

**SUDAAN**

```
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;
subgroup boy_r;
levels 2;
model pvtpct1c=agew1 boy_r hr_watch;
print /betafmt=f10.6 sebetafmt=f10.6;
run;
```

**SAS 9.1 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 9.1 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 3. Multilevel Models**

Data for 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 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 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 a MLM with a level for the school and a level for the adolescent. The algebraic formulas describing the model and assumptions appear below.

*Student-level model (Within or Level 1):*

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

where:

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

*School-level Model (Between or Level 2):*

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

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

where:

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

In this example, we will adjust for the sample design by using the sampling weights to adjust for unequal probability of selection. We use two different methods of scaling the sampling weights for estimating this model.

We followed PWIGLS Method 2 to scale the level 1 weight for the MLM analysis (Pfefferman, 1998) using gllamm. 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:

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

MLWIN and LISREL will automatically do this scaling for the user. In MLWIN, the weights are assumed to be independent of random effects.

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 model were 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 is just the product of the PWIGLS scaled level 1 weight and the level 2 weight. The analyst must create this weight for MPLUS.

Users of the Add Health data can download SAS and or Stata programs to help in doing the needed scaling of the weights. See appendix A.

The results of the estimation using each package are given in Table 4.3. 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.

Table 4.3. Results from estimation of 2-level model estimated with sampling weights.

| Parameter in<br>2-Level Model | MPLUS 4.0<br>Estimate (S.E) | LISREL 8.8<br>Estimate (S.E.) | MLWIN 2.02<br>Estimate (S.E.) | GLLAMM<br>Estimate (S.E.) |
|---|---|---|---|---|
| *Weights used* | MPML<br>Method A | PWIGLS<br>Method 2 | PWIGLS<br>Method 2 | PWIGLS<br>Method 2 |
| *Fixed Effects* | | | | |
| $\gamma_{00}$ (Intercept for $\beta_{0j}$) | 60.22 (1.09) | 59.26 (0.83) | 60.28 (1.17) | 60.22 (1.10) |
| $\gamma_{01}$ (Slope for $\beta_{0j}$) | -5.48 (1.49) | -3.01 (1.13) | -5.62 (1.65) | -5.48 (1.50) |
| $\gamma_{10}$ (Intercept for $\beta_{1j}$) | 0.032 (0.022) | 0.043 (0.022) | 0.030 (0.023) | 0.032 (0.022) |

|  |  |  |  |  |
|---|---|---|---|---|
| $\gamma_{11}$ (Slope for $\beta_{1j}$ ) | 0.13 (0.031) | 0.11 (0.028) | 0.130 (0.032) | 0.13 (0.031) |
| *Random Effects* |  |  |  |  |
| $\sigma^2_{\delta 0}$ (Var ($\delta_{0j}$ )) | 19.13 (6.94) | 9.16 (1.74) | 20.18 (6.04) | 19.32 (6.97) |
| $\sigma^2_{\delta 1}$ (Var ($\delta_{1j}$)) | 0.003 (0.002) | 0.001 (0.001) | 0.003 (0.001) | 0.003 (0.002) |
| $\sigma_{12}$ (Cov ($\delta_{0j}$,$\delta_{1j}$)) | -0.081 (0.097) | -0.063 (0.034) | -0.091 (0.071) | -0.079 (0.097) |
| $\sigma^2$ (Var ($e_{ij}$)) | 788.79 (16.96) | 798.15 (76.05) | 786.37 (86.62) | 788.81 (17.02) |

The program syntax used to compute the results in table 4.3 is given in table 4.4.

Table 4.4 Program syntax for multilevel analysis.

MULTILEVEL ANALYSIS PROGRAM STATEMENTS

```
MPLUS 4.0

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;
```

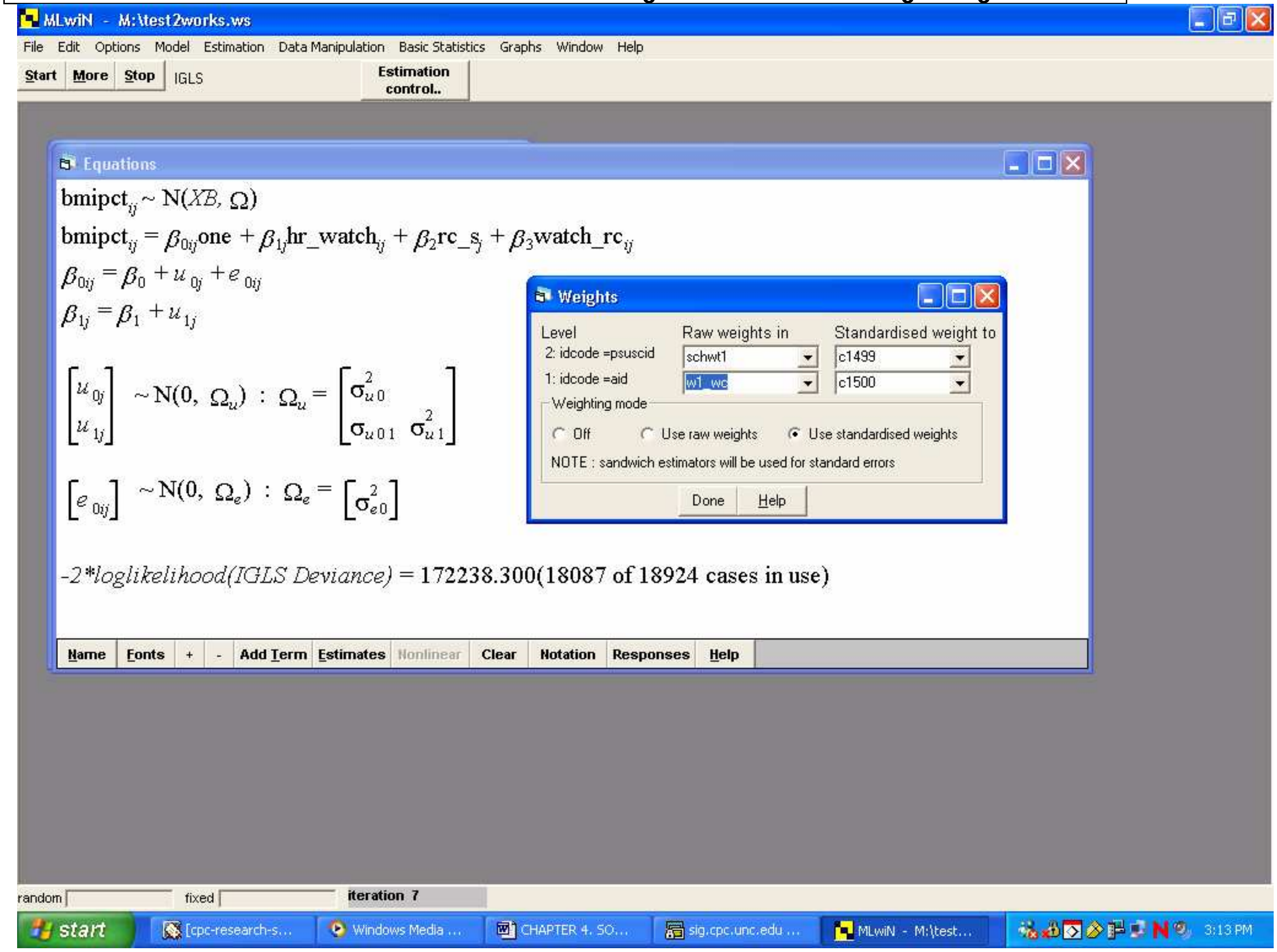
```
LISREL
OPTIONS OLS=YES CONVERGE=0.001000 MAXITER=10 COVBW=YES
OUTPUT=STANDARD ;
 TITLE=test;
 MISSING_DAT =-9999.000000 ;
 MISSING_DEP =-9999.000000 ;
 SY='M:\ls2lev4.psf';
 ID2=psuscid;
 WEIGHT2=schwt_1;
 WEIGHT1=w1_wc;
 RESPONSE=bmipct;
 FIXED=intcept hr_watch rc_s watch_rc;
 RANDOM1=intcept;
 RANDOM2=intcept watch_rc;
```

```
GLLAMM (in Stata 9)

generate mlwt2=schwt1
generate mlwt1=pw2r_w1
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(pw2_wt) trace adapt iter(20) nip(12)
```

**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.**



$$\text{bmipct}_{ij} \sim \text{N}(XB, \Omega)$$

$$\text{bmipct}_{ij} = \beta_{0ij}\text{one} + \beta_{1j}\text{hr\_watch}_{ij} + \beta_{2}\text{rc\_s}_{j} + \beta_{3}\text{watch\_rc}_{ij}$$

$$\beta_{0ij} = \beta_0 + u_{0j} + e_{0ij}$$

$$\beta_{1j} = \beta_1 + u_{1j}$$

$$\begin{bmatrix} u_{0j} \\ u_{1j} \end{bmatrix} \sim \text{N}(0,\ \Omega_u)\ :\ \Omega_u = \begin{bmatrix} \sigma^2_{u0} & \\ \sigma_{u01} & \sigma^2_{u1} \end{bmatrix}$$

$$\begin{bmatrix} e_{0ij} \end{bmatrix} \sim \text{N}(0,\ \Omega_e)\ :\ \Omega_e = \begin{bmatrix} \sigma^2_{e0} \end{bmatrix}$$

*-2*loglikelihood(IGLS Deviance)* = 172238.300(18087 of 18924 cases in use)

| 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

# Appendix A. Scaling weights for Multilevel Analysis

Stata and SAS programs for constructing sampling weights for estimating two-level models that can be used with several popular multilevel software packages can be downloaded from our website:

http://www.cpc.unc.edu/restools/data_analysis/ml_sampling_weights

Documentation is available from that website that provides information on using these programs to create the two-level weights, provides information about several popular multilevel software packages that allow these sampling weights to be used in estimation, and instruct the analyst in downloading and running these programs.

Both LISREL and gllamm will automatically do the scaling, so users of these software programs only need to specify the level 1 and level 2 weight components available with the Add Health data. Uses of gllamm and Mplus 4.1 and earlier will need to have the weights scaled as in described in Example 3 on multilevel models. Users of these programs can scale the weights by writing their own program or by using the SAS and Stata programs that we provide. The actual 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. 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 needed by gllamm.

Users of MPLUS 4.1 can just use the PWIGLS macro and multiple the level 2 weight and PWIGLS scaled level 1 weight together and get the needed combined weight. For this example, the MPLUS combined weight could be calculated as:

```
mp_wt_w1 = pw2r_w1*schwt1
```

Alternately, users can download the MPML_WT programs that will scale the weights according to the instructions given in Example 3.

Table A2. Example code used to construct composite weight for MPLUS used in example 3.

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 stata mpml_wt will the weight mp_wt_w1 for use in estimating 2-level models in Stata .

## Additional Information

Websites

Add Health: http://ww.cpc.unc.edu/addhealth

Center for Multilevel Modeling: http://multilevel.ioe.ac.uk/index.html

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

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

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

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

List Servers

Add Health to interact with other analysts: Send email to listproc@listserv.oit.unc.edu and in the body of the message type: `subscribe addhealth2` *`firstname lastname`*

Add Health to receive notification about data and documentation: Send email to listproc@listserv.oit.unc.edu and in the body of the message type: `subscribe addhealth` *`firstname lastname`*

Lists about survey software packages: http://www.stattransfer.com/lists.html

## References

Asparouhov, Timir, “Weighting for Unequal Probability of Selection in Latent Variable Modeling”, Webnote 7, http://www.statmodel.com/mplus/examples/webnotes/MplusNote71.pdf.

Asparouhov, Timir, “Weighting for Unequal Probability of Selection in Multilevel Modeling” WebNote 8, http://www.statmodel.com/mplus/examples/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, editors-in-chief Peter Armitage and Theodore Colton, John Wiley, New York, Volume 5, 1998, pages 4167-4174.

Chantala, K. and 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, S. B., “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", Internet edition http://www.arnoldpublishers.com/support/goldstein.htm

Levy, P. S., and Lemeshow, S., “Sampling of Populations Methods and Applications,” John Wiley & Sons, 1999. 525 p.

Littell, R. C., Milliken, G. A. Stroup, W. W., and Wolfinger, R. D., SAS System for Mixed Models", Cary, NC, SAS Institute, 1996.

Muthén, L. and 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, B. V., Barnwell, B. G., and Bieler, G. S., “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" http://gseweb.harvard.edu/~faculty/singer/.

Stapleton, “The Incorporation of Sample Weighs Into Multilevel Structural Equation Models”, Structural Equation Modeling, 9(4), 475-502.

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

Touraneau, R. and Hee-Choon, S. “National Longitudinal Study of Adolescent Health Grand Sample Weight,” Carolina Population Center, University of North Carolina at Chapel Hill

Williams, R. L., "A Note on Robust Variance Estimation for Cluster-Correlated Data", Biometrics 56, 645-646, June 2000.