Appropriate Analysis in Add Health:

Correcting for Design Effects & Selecting Weights

Ping Chen, PhD

Senior Research Scientist

Carolina Population Center

University of North Carolina at Chapel Hill

July 24, 2018
Bethesda, Maryland

UNC
CAROLINA
POPULATION
CENTER

Add Health
The National Longitudinal Study of Adolescent to Adult Health

Things to Cover - WHY and HOW

- Special features of Add Health design
- Choosing the correct sampling weight for analysis
- Common errors to avoid
- Preparing data for analysis
- Examples
- Subpopulation analysis
- Multilevel analysis

Special Features of Add Health Design

- Add Health is a **nationally representative sample** of adolescents in grades 7 through 12 between 1994 and 1995. It is a longitudinal cohort study that follows individuals from Wave I to Wave V (1995-2018).

Wave	Year	Age Range	Mean Age
I	1994 – 1995	11 – 21	15
II	1996	12 – 22	16
III	2001 – 2002	18 – 28	21
IV	2008 – 2009	24 – 34	28
V (Sample 1; N = 3,842)	2016 - 2017	32 - 42	36

UNC
CAROLINA
POPULATION
CENTER

Add Health
The National Longitudinal Study of Adolescent to Adult Health

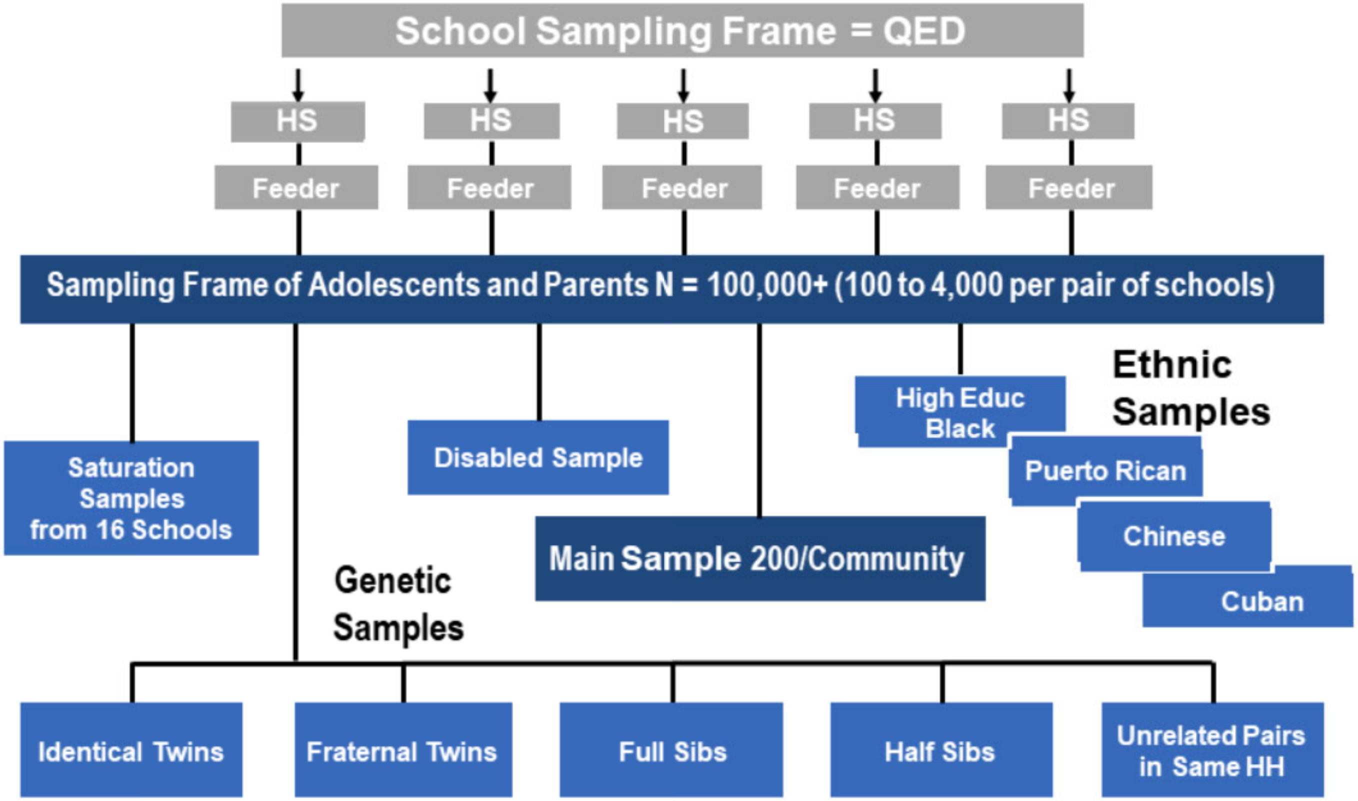
Special Features of Add Health Design

- First, the strategy of **multistage sampling** was used. This resulted in **clustered observations**.

- Second, it is a **probability-based** survey. However, like many other national studies, it is not a simple random sample. Each individual does **not** have an **equal probability of selection**. Oversampling of certain subgroups was used.

- Third, there was **stratification** in sampling.

Sampling Structure



School Sampling Frame = QED
HS
HS
HS
HS
HS
Feeder
Feeder
Feeder
Feeder
Feeder
Sampling Frame of Adolescents and Parents N = 100,000+ (100 to 4,000 per pair of schools)
Saturation Samples from 16 Schools
Disabled Sample
Main Sample 200/Community
High Educ Black
Ethnic Samples
Puerto Rican
Chinese
Cuban
Genetic Samples
Identical Twins
Fraternal Twins
Full Sibs
Half Sibs
Unrelated Pairs in Same HH

Source: http://www.cpc.unc.edu/projects/addhealth/design/researchdesign_3618_regular.pdf

Special Features of Add Health Design

- Multistage sampling and unequal probability of selection:

A list of 26,666 U.S. high schools was sorted based on enrollment size, school type, region, location, and percent white and then divided into groups for sampling.

With the unequal probability of selection, a sample of 132 schools was chosen.

School became the cluster identifier or **primary sampling unit** (PSU).

Special Features of Add Health Design

- From the 1994-1995 enrollment rosters for the selected schools, adolescents were chosen for an in-home interview with **unequal probabilities of selection**.
- It included a self-weighted core sample (roughly equal-sized samples selected from most schools).
- Oversampled subgroups, including high SES black, Cuban, Puerto Rican, Chinese adolescents, disabled youth, and a genetic sample (mono & dizygotic twins, full siblings, half siblings, and unrelated individuals in the same household).
- 16 purposively selected "saturated" schools where all students were recruited for in-home interviews.

Special Features of Add Health Design

- Stratification in sampling.

The Add Health sampling plan did not include a stratification variable. However, a **poststratification adjustment** was made to the sample weights so that region of country (variable REGION) could be used as a post-stratification variable.

Effects of Add Health Survey Design

- **Unequal probability of sample selection**
- **Clustering of individuals within schools**
  - Adolescents within the same school are not independent of one another.
  - Adolescent outcomes will be clustered, more similar, within schools than across schools.
- **Stratification by region**
- If these aspects of complex survey data are **ignored, point estimates and standard errors** can be **biased**, which may lead to incorrect inferences.

Panels of Data Affected

- Wave I School administrator data
- Wave I in-school survey
- Wave I, II, III, IV & V survey data

More Details about Add Health Design

Harris, M. Kathleen. 2013. “The Add Health Study: Design and Accomplishment.”

http://www.cpc.unc.edu/projects/addhealth/documentation/guides/DesignPaperWave_IIV.pdf

Tourangeau and Shin.1999. “Grand Sample Weight.”
http://www.cpc.unc.edu/projects/addhealth/data/guides/weights.pdf

Things to Cover

- Special features of Add Health design
- Choosing the correct sampling weight for analysis
- Preparing data for analysis
- Common errors to avoid
- Examples
    - Subpopulation analysis
    - Multilevel analysis

Correcting for Design Effects
Choosing the Correct Sampling Weight

- When researchers omit sample weights from the analysis of complex survey data, parameter estimates can be biased and incorrect inferences may be drawn.

- When sample weights are not used, findings cannot be generalized to the larger population of interest. Instead they only refer to the sample.

Available Weights in Add Health

Grand Sample Weight for Population-Average (single-level) Models	School-Level and Individual-Level for Multilevel Models
Cross-Sectional Weights	Cross-Sectional Weights
Longitudinal Weights	Longitudinal Weights

Cross-Sectional Grand Sample Weights
Single-Level (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	Adolescents in 1995 enrolled in grades 7-12 during the 1994-1995 academic year
Wave II (1996)	GSWGT2 (N=13,570)	Wave I respondents who were interviewed at Wave II	Adolescents in 1996 enrolled in grades 7-12 during the 1994-1995 academic year
Wave III (2001)	GSWGT3_2 (N=14,322)	Wave I respondents who were interviewed at Wave III	Adults in 2001 enrolled in grades 7-12 during the 1994-1995 academic year
Wave IV (2008)	GSWGT4_2 (N=14,800)	Wave I respondents who were interviewed at Wave IV	Adults in 2008 enrolled in grades 7-12 during the 1994-1995 academic year
Wave V Sample 1 (2016-2017)	GSWGT5_2 (N=3,704)	Wave I respondents who were interviewed at Wave V Sample 1	Adults in 2016-17 enrolled in grades 7-12 during the 1994-1995 academic year

Choosing a Sampling Weight for Analysis
Cross-Sectional Weights

- Research questions tend to investigate association rather than causation.

- Scenario #1: both predictor and outcome variables are collected at the same point in time (from the same Wave, either Wave I, II, III, or IV).

- Scenario #2: the **outcome variable** is **from one wave of data**, either Wave I, II, III or IV, but predictors (or covariates) are from previous wave(s) or a combination of multiple waves. Under this circumstance, choose the cross-sectional weight for the wave at which the outcome was measured.

Using Cross-Sectional Weights

X Variable	Y Variable	Weight
Same wave as Y (e.g. Wave IV education)	One Wave (e.g. WIV BMI)	Cross-sectional (population-average model) (e.g. gswgt4_2)
Multiple Waves (e.g. WI GPA, WIII edu, WIV physical activity)	One Wave (e.g. WIV BMI)	Cross-sectional (population-average model) (e.g. gswgt4_2)

Longitudinal Weights
Single-Level (Population Average) Models

Data Set (Year collected)	Sampling Weight Variable (N)	Sample	Target Population
Wave III (2001)	GSWGT3 (N=10,828)	Eligible Wave I Respondents who were interviewed at both Wave II & Wave III	Adolescents enrolled in US schools during the 1994-1995 academic year for the grades 7-12 in 1994-1995
Wave IV (2008)	GSWGT4 (N=9,421)	Eligible Wave I respondents who were interviewed at Wave II, III & IV	Adolescents enrolled in grades 7-11 during 1994-1995 interviewed in 1995, 1996, 2001 & 2008
Wave IV (2008)	GSWGT134 (N=12,288)	Eligible Wave I respondents who were interviewed at Wave III & IV	Adolescents enrolled in grades 7-11 during 1994-1995 interviewed in 1995, 1996 & 2008
Wave V (2016-2017)	GSWGT145 (N=3,381)	Eligible Wave I respondents who were interviewed at Wave IV & V	Adolescents enrolled in grades 7-11 during 1994-1995 interviewed in 1995, 2008 & 2017

Choosing a Sampling Weight for Analysis
Longitudinal Weights

- Longitudinal analysis is used to investigate **changes in measurements** taken on respondents over time

- **The outcome variable is measured multiple times**

  (*Note: if the covariates are from multiple waves but the outcome variable is from one wave, longitudinal weights CANNOT be used*.)

Using Longitudinal Weights

X Variable	Y Variable	Weight
Wave I (e.g. WI GPA)	Multiple Waves for a sample of respondents who have data at *every* wave (e.g. WI, II, III, IV BMI)	Longitudinal (e.g. GSWGT4 in GEE estimation)

Sampling Weights for Wave III Special Sub-Samples for Estimating Single-Level (population average) Models

Data Set (Year collected)	Sampling Weight Variable (N)	Sample	Represented Population of Interest
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 of Adolescents enrolled in Grades 7-12 in 1994-1995
	TWGT3_2 (N=11,637)	Wave III Education Sample: Eligible Wave I respondents interviewed at Wave III.	Adolescents enrolled in Grades 7-12 in 1994-1995 involving in high school transcripts
	TWGT3 (N=8,847)	Wave III Education Sample: Eligible Wave II respondents interviewed at Wave III.	Same as above

Sampling Weights for Wave III Special Sub-Samples for Estimating Single-Level (population average) Models

Data Set (Year collected)	Sampling Weight Variable (N)	Sample	Represented Population of Interest
Wave III (2001)	MGENCRWT (N=14,322) (MGEN Cross-Sectional Weight)	MGEN Sample: special sample selected for testing urine for mycoplasma genitalium at Wave III.	Adolescents enrolled in Grades 7-11 during 1994-1995 and interviewed in 1995 & 2001 (Involving in MGEN sample)
	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.	Adolescents enrolled in Grades 7-11 during 1994-1995 and interviewed in 1995, 1996, & 2001 (Involving in MGEN sample)
	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 involving in HPV sample
	HPLORWT (N=4,945) (HPV Cross-Sectional 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 in HPV sample

** GWAS sample weights for GWAS data (latest release)

Choosing a Sampling Weight for Analysis
Multilevel Model

- 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.

- Add Health makes two levels of weight components available to users. The level 1 weight component refers to individuals (respondents) and level 2 weight component refers to PSU (schools).

- Note that the two-level sampling weights need to be scaled before you can run a Multilevel model in different packages. Scaling methods may differ depending on which package you use.

- There are two different methods for scaling the sampling weights: PWIGLS METHOD 2 and MPML METHOD A.

UNC CAROLINA POPULATION CENTER

Add Health The National Longitudinal Study of Adolescent to Adult Health

Cross-Sectional Weight Components for Multilevel Model

Interview (Year Collected)	Level 2 Weight Component (N)	Level 1 Weight Component (N)	Sample	Target Population
In-School (1994)	SCHWT128 (N=128)	INSCH_WC (N=83,135)	Adolescents chosen with a known probability of being selected from 1994-1995 enrollment rosters of US schools.	Adolescents in 1995 enrolled in grades 7-12 during 1994-1995
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	Adolescents in 1995 enrolled in grades 7-12 during the 1994-1995 academic year
Wave II (1996)	SCHWT1 (N=132)	W2_WC (N=13,568)	Wave I respondents who were interviewed at Wave II.	Adolescents in 1996 enrolled in grades 7-12 during the 1994-1995 academic year
Wave III (2001)	SCHWT1 (N=132)	W3_2_WC (N=14,322)	Wave I respondents who were interviewed at Wave III.	Adults in 2001 enrolled in grades 7-12 during the 1994-1995 academic year
Wave IV (2008)	SCHWT1 (N=132)	W4_2_WC (N=14,800)	Wave I respondents who were interviewed at Wave IV.	Adults in 2008 enrolled in grades 7-12 during the 1994-1995 academic year

UNC
CAROLINA
POPULATION
CENTER

Add Health
The National Longitudinal Study of Adolescent to Adult Health

Using Cross-Sectional Two-Level Weights

X Variable	Y Variable	Weight
Same wave as Y (e.g. School-level recreation center; Individual-level Wave I parental education)	One Wave (e.g. WIV BMI)	Two-level Cross-sectional (e.g. SCHWT1; W4_2_WC)
Multiple Waves (e.g. School-level recreation center; Individual-level WI GPA, WIII edu, WIV household income)	One Wave (e.g. WIV BMI)	Two-level Cross-sectional (e.g. SCHWT1; W4_2_WC)

Longitudinal Weight Components for Multilevel Model

Interview (Year Collected)	Level 2 Weight Component (N)	Level 1 Weight Component (N)	Sample	Represented Population
Wave III (2001)	SCHWT1 (N=132)	W3_WC (N=10,828)	Eligible Wave I Respondents who were interviewed at both Wave II & Wave III.	Adolescents in 2001 enrolled in grades 7-12 during 1994-1995
Wave IV (2008)	SCHWT1 (N=132)	W4_WC (N=9,421)	Eligible Wave I respondents who were interviewed at Wave II, III & IV.	Adolescents in 2008 enrolled in grades 7-12 during 1994-1995

Using Cross-Sectional Two-Level Weights

X Variable	Y Variable	Weight
Same Wave as Y (e.g. School-level recreation center; Individual-level Wave I parental education)	Multiple Waves for a sample of respondents who have data at *every* wave (e.g. WI, II, III & IV BMI)	Two-level Longitudinal (e.g. SCHWT1; W4_WC)
Multiple Waves (e.g. School-level recreation center; Individual-level WI GPA, WIII edu, WIV household income)	Multiple Waves (e.g. WI, II, III & IV BMI)	Two-level Cross-sectional (e.g. SCHWT1; W4_WC)

Choosing a Sampling Weight for Analysis
Time-to-Event (Survival) Analysis

- When individuals are observed over time and the outcome is the occurrence and timing of a specific event, a survival analysis is appropriate. Examples of events: death, onset of a disease, first pregnancy, or first marriage.

- The event may not be observed for all respondents.

- Choice of sampling weight will usually be determined by the data collected at the **earliest** time point.

Sampling Weights for Time-to-Event (Survival) Analysis
One-Time Point

Data Source (Y from One Wave)	**Weight for Population Average Models**	**Weights for Multilevel Models**	**Sample**	**Target Population**
Wave I only (1995)	GSWGT1 (N=18,924)	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.	Adolescents in 1995 enrolled in grades 7-12 during 1994-1995
Wave II only (1996)	GSWGT2 (N=13,570)	W2_WC (N=13,568)	Wave I respondents who were interviewed at Wave II.	Adolescents in 1996 enrolled in grades 7-12 during 1994-1995
Wave III only (2001)	GSWGT3_2 (N=14,322)	W3_2_WC (N=14,322)	Wave I respondents who were interviewed at Wave III.	Adults in 2001 enrolled in grades 7-12 during 1994-1995
Wave IV only (2008)	GSWGT4_2 (N=14,800)	W4_2_WC (N=14,800)	Wave I respondents who were interviewed at Wave IV.	Adults in 2008 enrolled in grades 7-12 during 1994-1995

UNC
CAROLINA
POPULATION
CENTER

Add Health
The National Longitudinal Study of Adolescent to Adult Health

Sampling Weights for Time-to-Event (Survival) Analysis
Multiple Time Points

Data Source (Y from Multiple Waves)	**Weight for Population Average Models**	**Weights for Multilevel Models**	**Target Population**
Wave I & II	GSWGT1 (N=18,924)	SCHWT1 (N=132) W1_WC (N=18,924)	Adolescents in 1995 enrolled in grades 7-12 during 1994-1995
Wave II & III	GSWGT2 (N=13,570)	SCHWT1 (N=132) W2_WC (N=13,568)	Adolescents in 1996 enrolled in grades 7-12 during 1994-1995
Wave I, II, & III	GSWGT1 (N=18,924)	SCHWT1 (N=132) W1_WC (N=18,924)	Adolescents in 1995 enrolled in grades 7-12 during 1994-1995
Wave I, II, III, & IV	GSWGT1 (N=18,924)	SCHWT1 (N=132) W1_WC (N=18,924)	Adolescents in 1995 enrolled in grades 7-12 during 1994-1995

UNC CAROLINA POPULATION CENTER

Add Health The National Longitudinal Study of Adolescent to Adult Health

Things to Cover

- Special features of Add Health design
- Choosing the correct sampling weight for analysis
- Common errors to avoid
- Preparing data for analysis
- Examples
- Subpopulation analysis
- Multilevel analysis

Avoiding Common Errors

- ***Do not ignore clustering, stratification, and unequal probability of selection when analyzing the Add Health data.*** Otherwise, biased estimates and possibly false-positive hypothesis test results may be generated. Point estimates (means, regression parameters, proportions, etc.) are only affected by the weights. Variance estimates are affected by the clustering, stratification, weight, and design type.

- ***Do not include respondents who are missing sampling weights in your analysis when your goal is to obtain national estimates.*** Otherwise, you will have the wrong sample size.

- ***Do not subset the probability sample (of adolescents who have weights) by deleting cases that are not in the subsample.*** Subsetting the data may cause an incorrect number of PSUs to be used in the variance computation formula. Most software packages for analyzing data from sample surveys provide special commands for conducting subpopulation analysis.

Avoiding Common Errors

- ***Do not use the Sampling Weight as a Frequency or Analytical Weight during your 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 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. For example, a sampling weight of 25 means that the data from the recruited individual is representative of 25 people in the population of interest.

Avoiding Common Errors

- 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 the analysis.

- The analyst needs to make sure that the Add Health weights are used as sampling weights.

Avoiding Common Errors

- ***Do not normalize 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 by the developers of the software or documentation supplied with the software. If you normalize the weights, estimates of population totals will be incorrect even if you use the survey software.

Things to Cover

- Special features of Add Health design
- Choosing the correct sampling weight for analysis
- Common errors to avoid
- Preparing data for analysis
- Examples
- Subpopulation analysis
- Multilevel analysis

Preparing Data for Analysis

- Determine the wave(s) of data you need for your analysis and construct desired variables.
- Exclude/delete cases that have missing in sampling weights.
- Identify the attributes & elements of the sample design (with replacement Design, strata variable, cluster variable, weight variable).

Design Type: Specify With Replacement as the Design Type

Even though schools were not placed back on the list before the next school was selected, 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 replacement sampling.

Stratum Variable: Use 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. This involved using the total number of schools on the sampling frame for each region (Northeast, Midwest, South, and West) of the country. For each region, an adjustment was made to the initial school weights so that the sum of the school weights was equal to the total number of schools on the sampling frame.

Preparing Data for Analysis

Cluster Variable or Primary Sampling Unit (PSU): Use the School Identifier (PSUSCID)

This variable is used for the In-School, Wave I, II, III, IV, and V sample 1 data. The sampling units in the Add Health study are middle and high schools from the United States, hence 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 conduct and choose the appropriate weight variable.

Note that region and psuscid are included in the same files as the weight variables.

Use subpopulation analysis if your data only include a subset of the original Add Health sample.

Things to Cover

- Special features of Add Health design
- Choosing the correct sampling weight for analysis
- Preparing data for analysis
- Common errors to avoid
- Examples
- Subpopulation analysis
- Multilevel analysis

Example 1. Descriptive Statistics

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

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

SAS syntax:

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

STATA syntax:

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

Example 1. Descriptive Statistics

Parameter estimates and standard errors to predict the average number of hours TV watched during a week by adolescents

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

Example 2. Regression Example for Single-Level Model

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

STATA syntax:

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

SAS syntax:

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

Example 2. Regression Example for Population-Average Models

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

Parameter	SAS Estimate (Std Err)	Stata 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)

Things to Cover

- Special features of Add Health design
- Choosing the correct sampling weight for analysis
- Preparing data for analysis
- Common errors to avoid
- Examples
- Subpopulation analysis
- Multilevel analysis

Subpopulation Analysis

- When using survey data, it is common that researchers want to analyze only a certain group of respondents, perhaps only a sub-sample of women, respondents 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. Using the subpopulation option is extremely important when analyzing survey data with a sub-sample.
- If the data set is a subset, which means that observations not included in the sub-sample/subpopulation are deleted from the data set, the standard errors of the estimates may 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 used in the calculation of the point estimate, but all cases are used in the calculation of the standard errors.

Subpopulation Analysis

- **Scenario I**

- Often some of the respondents did not answer the questions that you use in your analysis. This means that the parameters will not be estimated from the full sample. In this case, you are actually analyzing a subset of the data. We recommend 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, since you will want the results from all models to be based on the same observations.

- Note if you have one or some variables that have a large number of missing, we recommend you conduct some imputation for missing values instead of using the subpopulation option.

Subpopulation Analysis

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

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

Subpopulation Analysis

- **Scenario II**
- 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 the subpopulation option to identify a sub-sample (subpop) of N=10,285.

- **Scenario III**
- Before you do the analysis, you often need to prepare a subpopulation indicator variable. Suppose you are interested in studying a subpop of Mexican Americans who reported a history of drug or alcohol use. You then 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.” Then you need to include this variable in the subpopulation option in your analysis.

Subpopulation Analysis

ID	race	drug_use	Alcohol_use	Weight (lbs)	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

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

Subpopulation Analysis – Example 1. 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)?

STATA INCORRECT way of subsetting data: Deleting cases that are not in the subpopulation to subset data

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

STATA 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: males (0) & females (1)

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

SAS 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 fm_wt;
domain female;
run;
```

		INCORRECT Deleting cases that are not in the subpopulation to subset data	CORRECT Subpopulation option in the software	Use DOMAIN statement to specify subpopulation
	Variable	Stata Estimate (Std Err)	Stata Estimate (Std Err)	SAS 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		---	---	---
Population size		10843943	---	---
Design DF		127	128	---
	hr_tv	14.55 (.41)	14.55 (.41)	14.55 (.41)

Subpopulation Analysis in Different Statistical Packages

- 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. But the SAS SURVEY procedure can be tricked into computing the correct variance and standard errors when analyzing subpopulations.

- ***Set weights outside the subpopulation to a very small value (close to zero).*** ***Note*** that SAS deletes observations that have a zero value for the sampling weight, so do not use zero value for weights.

- Stata has a subpopulation option.

Subpopulation Analysis – Example 2. Multivariate Analysis

Research Question: What is the effect of watching TV on PVT scores for adolescents attending RURAL schools (data from Wave I in-home questionnaire)? (The variable rural is coded as 1= rural school, 0=non-rural school.)

STATA 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;
```

UNC CAROLINA POPULATION CENTER

Add Health
The National Longitudinal Study of Adolescent to Adult Health

Subpopulation Analysis – Example 2. Multivariate Analysis

Research Question: What is the effect of watching TV on the PVT scores for adolescents attending RURAL schools (data from Wave I in-home questionnaire)? (The variable rural is coded as 1= rural school, 0=non-rural school.)

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;
```

Note: this is a no-intercept model.

Subpopulation Technique	INCORRECT Subset Data	CORRECT Subpopulation option in software	Set Weights outside subpopulation to 0.00001	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)

* Results in red refer to estimated standard errors when the method of listwise deletion (e.g. with/without “if” statement in Stata) was used . They are different/biased from the ones estimated by the subpopulation analysis.

Things to Cover

- Special features of Add Health design
- Choosing the correct sampling weight for analysis
- Preparing data for analysis
- Common errors to avoid
- Examples
- Subpopulation analysis
- Multilevel analysis

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.

- Thus Add Health makes two levels of weight components available to users. The level 1 weight component pertains to individuals (respondents) and level 2 weight pertains to PSU (schools).

Scaling Sampling Weights

- Note that the two-level sampling weights need to be scaled before you run a multilevel model in different packages. Scaling methods may differ depending on which package you use.
- There are two different methods of scaling the sampling weights for estimating this type of model.

PWIGLS METHOD 2

- One 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 or procedures that use this PWIGLS Method2 scaling method, including XTMIXED in Stata, GLLAMM in Stata, MLWIN, and LISREL.

Scaling Sampling Weights

MPML METHOD A

- Another scaling method is called 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 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:

$$mp\_wt\_w1_{i,j} = \frac{w1\_wc_{i|j} * schwt1_{j}}{\left(\dfrac{\sum_{i}^{n_1} w1\_wc_{i|j}}{n_j} \right)}$$

- This is just the product of the PWIGLS scaled level 1 weight and level 2 weight. The analyst can use the user written program, MPML_WT, to create this weight for MPLUS.

A summary of Scaling Methods based on Features of Different Statistical Packages/Procedures to Run a Multilevel Model

	Use PWIGLS Method 2	Need to use PWIGLS program to do the scaling before running the Multilevel Model	Use MPML Method A	Need to use MPML_WT program to do the scaling before running the Multilevel Model
MIXED 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
MPlus	No	NA	Yes	Yes

Note: **Users of the Add Health data can download SAS and/or Stata programs, PWIGLS and/or MPML_WT to scale the two-level weight component variables. See appendix A in the Guidelines.**

UNC
CAROLINA
POPULATION
CENTER

Add Health
The National Longitudinal Study of Adolescent to Adult Health

Program Syntax for Multilevel Analysis – Similar Data Set

MIXED (in Stata 14 and 15)

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

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

Program Syntax for Multilevel Analysis

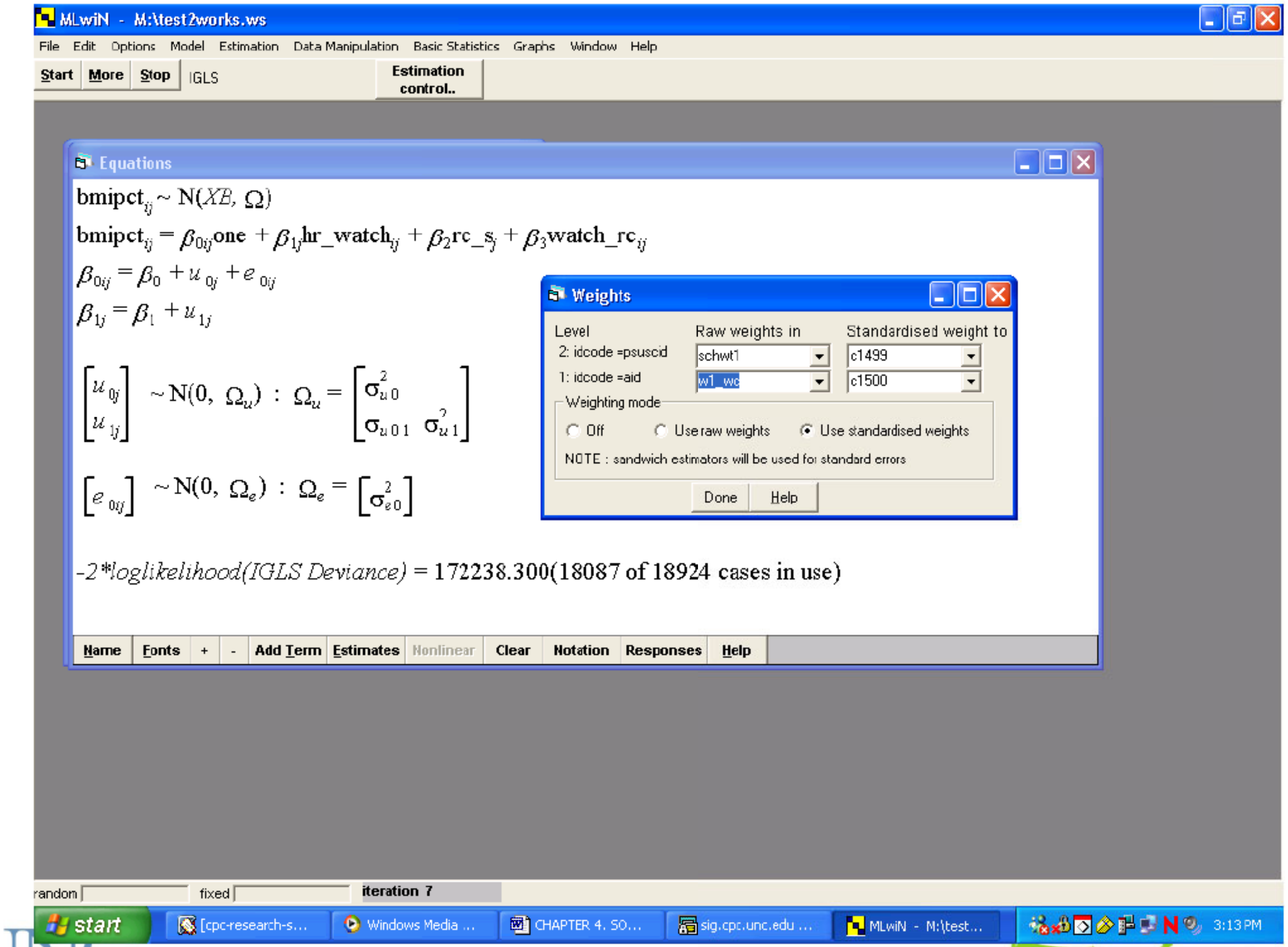
LISREL

*** 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=schwt1;
 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
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

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

$$\text{bmipct}_{ij} = \beta_{0ij}\text{one} + \beta_{1j}\text{hr\_watch}_{ij} + \beta_2\text{rc\_s}_j + \beta_3\text{watch\_rc}_{ij}$$

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

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

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

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

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

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

Example Code Used to Construct Weights for gllamm Cross-Sectional Analysis – PWIGLS Method 2

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);
```

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)
```

Scaling Weights for Multilevel Cross-Sectional Analysis

- Users of MPLUS can just use the PWIGLS macro and multiply the level 2 weight and PWIGLS scaled level 1 weight together to 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.

Example Code Used to Construct Weights for Mplus Cross-Sectional Analysis -- **MPML METHOD A**

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)
```

Scaling Weights for Multilevel Cross-Sectional Analysis

- 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.

Scaling Weights for Multilevel Analysis

- The variables psuscid (identifying the school), level 2 weight component (schwt1), respondent identifier (aid), and 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.

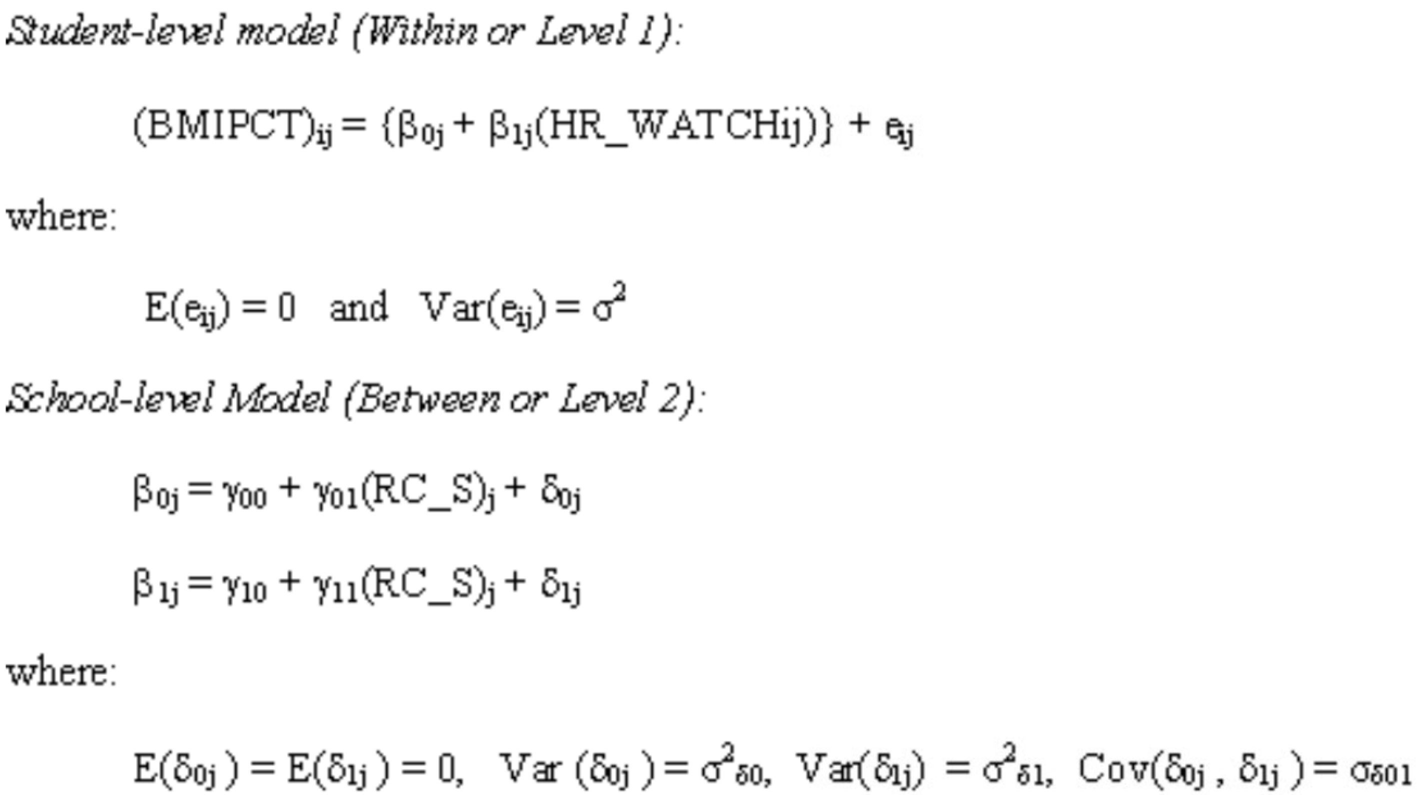
Example for Two-Level Linear Mixed-Effects Model

Variable List	Variable Definition	Variable Name
School-level X variable @ WI	the availability of an on-site school recreation center	RC_S
Individual-level X variable @ WI	hours watching TV or playing video or computer games during the past week.	HR_WATCH
Y @ WI	percentile body mass index	BMIPCT

Example of Two-Level Linear Mixed-Effects Model

- Data for this example illustrating the multilevel software packages come from the School Administrator Survey and the Wave I In-home Survey.
- Outcome variable is percentile body mass index (BMIPCT) .
- Student-level independent variable: hours watching TV or playing video or computer games during the past week (HR_WATCH).
- School-level independent variable: the availability of an on-site school recreation center (variable RC_S).

Example of Two-Level Linear Mixed-Effects Model



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:

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

Program Syntax for Multilevel Analysis

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.

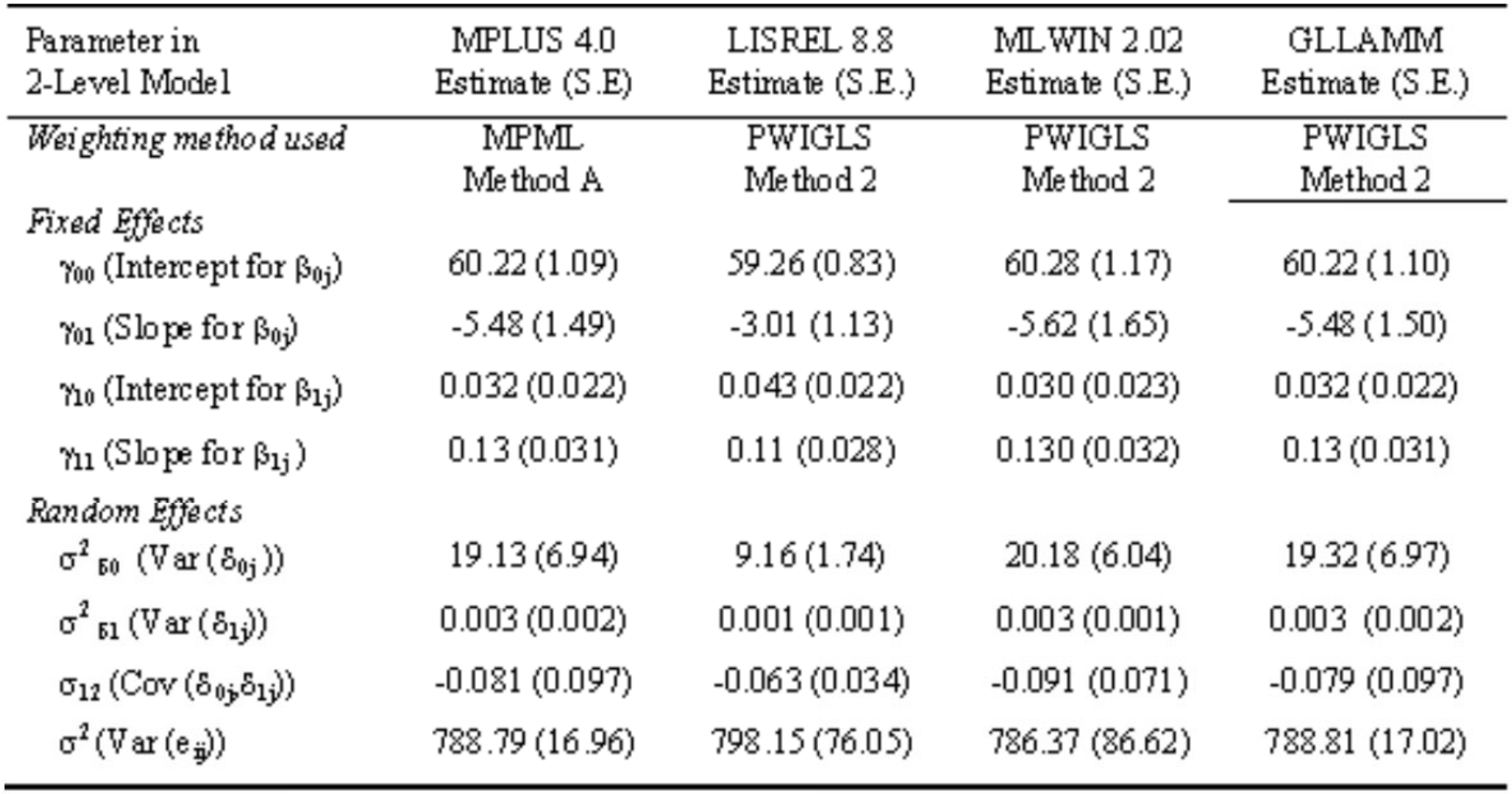
```
generate mlwt2=schwt1
generate mlwt1=pw2r_w1
generate one=1
eq sch_int: one
eq sch_slop: hr_watch
gllamm bmipct rc_s hr_watch watch_rc , i(sch_id) nrf(2) ///
     eqs(sch_int sch_slop) pweight(mlwt) trace adapt iter(20) nip(12)
```

Program Syntax for Multilevel Analysis

MPLUS 4.0

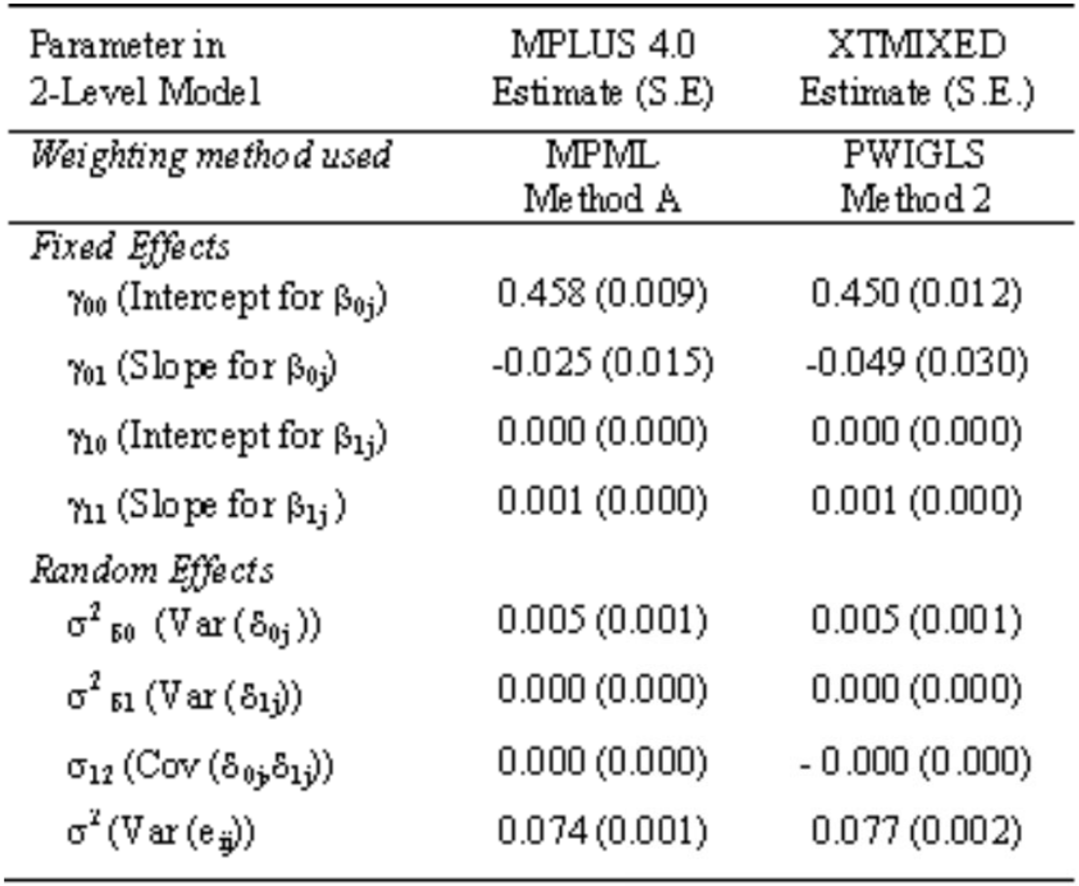
```
*** First, use MPML_WT program to scale the weights (see Appendix A):

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;
```

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)

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)

Note: Variables in this table are defined differently from the ones in the previous table, so the results here are different from the ones in the previous table.

UNC
CAROLINA
POPULATION
CENTER

Add Health
The National Longitudinal Study of Adolescent to Adult Health

Multilevel Mixed-Effects Models in Stata14

```
svyset psuscid, weight(schwt1) strata(region) || aid, weight(w1_wc)
```

Individual-level weight component variable

mecloglog	Multilevel mixed-effects complementary log-log regression
meglm	Multilevel mixed-effects generalized linear model
meintreg	Multilevel mixed-effects interval regression
melogit	Multilevel mixed-effects logistic regression
Menbreg	Multilevel mixed-effects negative binomial regression
Meologit	Multilevel mixed-effects ordered logistic regression
Meoprobit	Multilevel mixed-effects ordered probit regression
Mepoisson	Multilevel mixed-effects poisson regression
Meprobit	Multilevel mixed-effects probit regression
Mestreg	Multilevel mixed-effects parametric survival models
Metobit	Multilevel mixed-effects tobit regression

Multilevel Mixed-Effects Logit Model

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 for schools		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

Multilevel Mixed-Effects Logit Model

```
svyset psuscid, weight(schwt1) strata(region) || aid, weight(w1_wc)
```

```
svy: melogit w1obese w1hr_tv w1schrecrctr || psuscid:
```

```
svyset psuscid, weight(schwt1) strata(region) || aid, weight(w1_wc)
```

```
svy, subpop(nonmiss): melogit w1obese w1hr_tv w1schrecrctr || psuscid:
```

UNC
CAROLINA
POPULATION
CENTER

Add Health
The National Longitudinal Study of Adolescent to Adult Health

Summary

- Add Health has special survey design features.
- Users need to account for those special features (clustering, stratification, and unequal probability of selection) when they analyze Add Health data. Otherwise, inferences drawn from the results may not be correct.
- Use cluster (psuscid), stratification (region), and weight variables if possible.
- Select the correct weight variable depending on the type of analysis you choose.
- Use the subpopulation option, if you are analyzing a subsample of Add Health data.
- Scale two-level weight component variables when analyzing a multilevel model with both school-level and individual-level data.

More Questions

- Add Health provides online documentation outlining guidelines about how to account for the design effects of Add Health data and use weights when conducting data analysis.

Chen, Ping and Kim Chantala. 2014. *“Guidelines for Analyzing Add Health Data.”*

http://www.cpc.unc.edu/projects/addhealth/documentation/guides/copy_of_wt_guidelines_20161213.pdf

- *Questions: addhealth@unc.edu*

UNC
CAROLINA
POPULATION
CENTER

Add Health
The National Longitudinal Study of Adolescent to Adult Health

Acknowledgment

This research uses data from Add Health, a program project directed by Kathleen Mullan Harris and designed by J. Richard Udry, Peter S. Bearman, and Kathleen Mullan Harris at the University of North Carolina at Chapel Hill, and funded by grant P01-HD31921 from the Eunice Kennedy Shriver National Institute of Child Health and Human Development, with cooperative funding from 23 other federal agencies and foundations. Special acknowledgment is due Ronald R. Rindfuss and Barbara Entwisle for assistance in the original design. Information on how to obtain the Add Health data files is available on the Add Health website (http://www.cpc.unc.edu/addhealth). No direct support was received from grant P01-HD31921 for this analysis.

UNC
CAROLINA
POPULATION
CENTER

Add Health
The National Longitudinal Study of Adolescent to Adult Health

THANK YOU

UNC
CAROLINA
POPULATION
CENTER

Add Health
The National Longitudinal Study of Adolescent to Adult Health