Appropriate Analysis in Add Health
Correcting for Design Effects & Selecting Weights

Ping Chen
Carolina Population Center
University of North Carolina at Chapel Hill

June 26, 2014
Bethesda, Maryland

UNC CAROLINA POPULATION CENTER

Add Health
The National Longitudinal Study of Adolescent to Adult Health

Things to Cover

- Special features of Add Health design
- How to choose the correct sampling weight for analysis
- Steps of Preparing Data for Analysis
- Examples
  - Descriptive Statistics
  - OLS regression
  - Subpopulation Analysis
  - Multilevel Model

Chen and Chantala. 2014 “Guidelines for Analyzing Add Health Data.”
http://www.cpc.unc.edu/projects/addhealth/data/guides/wt-guidelines.pdf

Things to Cover

- Special features of Add Health design
- How to choose the correct sampling weight for analysis
- Steps of Preparing Data for Analysis
- Examples
  - Descriptive Statistics
  - OLS regression
  - Subpopulation Analysis
  - Multilevel Model

Special Features of Add Health Design

- Add Health is a national representative sample of adolescents in grades 7 through 12 in the United States in 1994-95. It is a longitudinal study that follows individuals through adolescence to the early adulthood with four waves of in-home interviews.
- 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.

Special Features of Add Health Design

- First, the strategy of multistage sampling was used. This resulted in clustered observations.
- Second, the probabilities of selection of the observations are not equal; oversampling of certain subgroups in population was employed.
- Third, stratification in sampling.

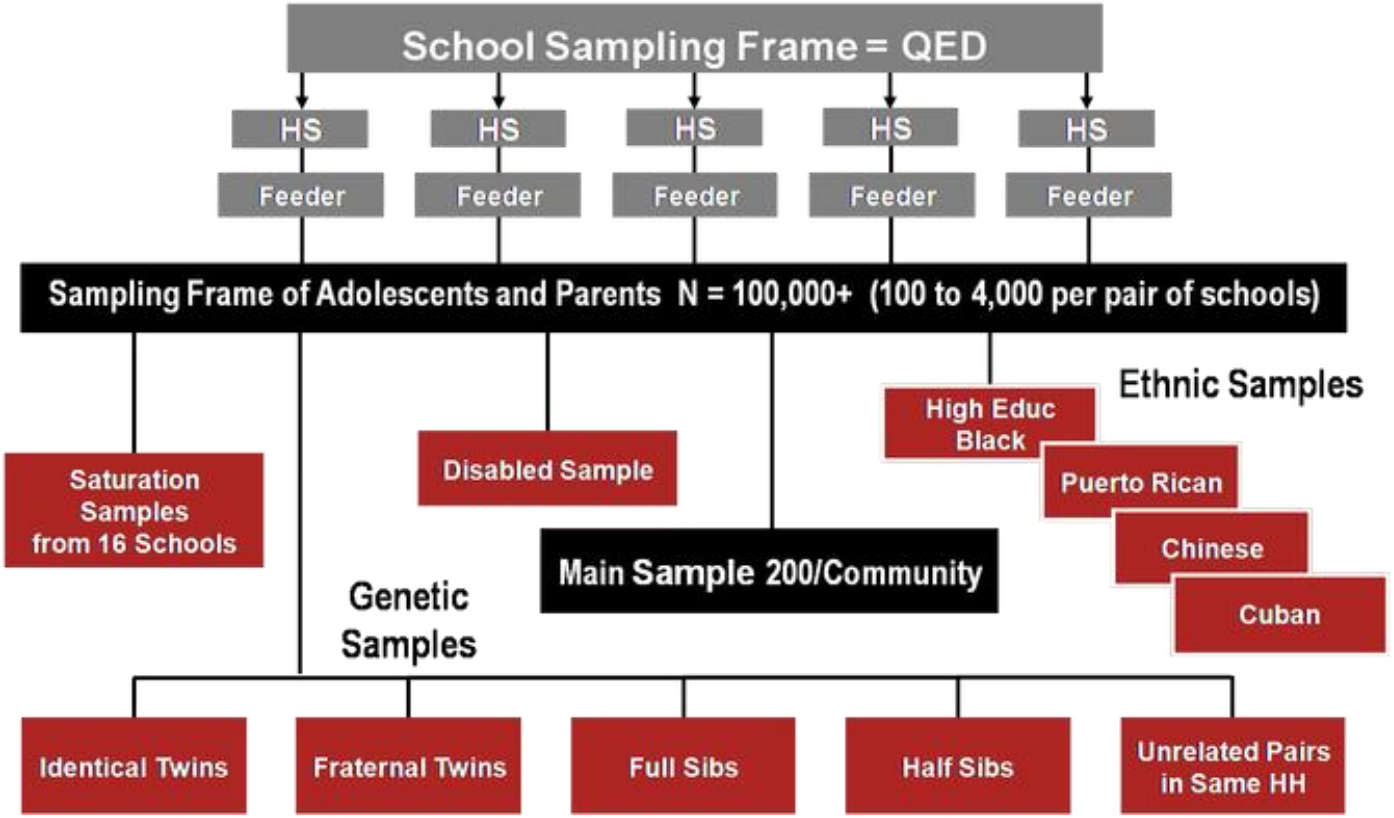
Special Features of Add Health Design

- Multistage sampling and unequal probability of selection:
- The sampling frame was derived from Quality Education Database (QED) comprised of 26,666 U.S. high schools.
- From this frame a stratified sample of 80 high schools (with an 11th grade and more than 30 students) with probability proportional to size.
- Schools were stratified by region, urbanicity, school type, ethnic mix and size.
- For each high school selected, we identified and recruited one of its feeder schools (typically a middle school) with probability proportional to its student contribution to the high school. A total of of 52 feeder (junior high or middle) schools were selected.
- Some schools spanned from grades 7 to 12, then no feeder schools were recruited.
- With this 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 and those not on the rosters that completed the in-school questionnaire, adolescents were chosen with unequal probabilities of selection.
- First, a core sample was drawn by stratifying students in each school by grade and sex and this yields a nationally representative core sample (roughly equal-sized samples selected from most schools) of 12,105 adolescents in grades 7 to 12.
- Then, Add Health drew supplemental samples, oversampling certain groups based on ethnicity (Cuban, Puerto Rican, and Chinese), genetic relatedness to siblings (with twins, full siblings, half siblings, and unrelated individuals in same household), adoption status, disabled youth, and black adolescents with highly educated parents. We also purposively selected 16 schools of all students included.

Sampling Structure



Special Features of Add Health Design

- First, the strategy of multistage sampling was used. This resulted in clustered observations.
- Second, the probabilities of selection of the observations are not equal; oversampling of certain subgroups in population was employed.
- Third, stratification in sampling.

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.

Special Features of Add Health Design

- First, the strategy of multistage sampling was used. This resulted in clustered observations.
- Second, the probabilities of selection of the observations are not equal; oversampling of certain subgroups in population was employed.
- Third, stratification in sampling.

Panels of Data Affected

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

More Details about Add Health Design

- Tourangeau and Shin.1999. “Grand Sample Weight.” http://www.cpc.unc.edu/projects/addhealth/data/guides/weights.pdf
- “Add Health Research Design Waves I-IV.” http://www.cpc.unc.edu/projects/addhealth/design/slideshow
- Harris. 2013. “The Add Health Study: Design and Accomplishments.” http://www.cpc.unc.edu/projects/addhealth/data/guides/DesignPaperWIIV.pdf

Effects of Add Health Survey Design

- In summary, Add Health survey design includes special features, including clustered observations as a result of multistage sampling, unequal probabilities of selection of the observations, and stratification

- If these aspects of complex survey data are ignored, point estimates and standard errors may be biased, hence potentially leading to incorrect inferences made by researchers.

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

Chen and Chantala. 2014 “Guidelines for Analyzing Add Health Data.”

http://www.cpc.unc.edu/projects/addhealth/data/guides/wt-guidelines.pdf

Things to Cover

- Special features of Add Health design
- How to choose the correct sampling weight for analysis
- Steps of Preparing Data for Analysis
- Examples
  - Descriptive Statistics
  - OLS regression
  - Subpopulation Analysis
  - Multilevel Model

Correcting for Design Effects
Choosing the Correct Sampling Weight

- When researchers omit sample weights from the analysis of complex survey data, parameter estimates are biased and incorrect inferences are drawn.
- In addition, when sample weights are not used, findings cannot be generalized to the larger population of interest. Instead, they can only be generalized to the sample.

Available Weights in Add Health

- Single-level cross-sectional weights
- Single-level longitudinal weights
- Multi-level cross-sectional weights
- Multi-level longitudinal weights
- Single-level cross-sectional and longitudinal weights for sub-samples

Choosing a Sampling Weight for Analysis
Cross-Sectional Analysis

- Research questions tend to investigate association rather than causation.
- One scenario: both predicting and outcome variables are collected at the same point in time (from the same Wave, either Wave I, II, III, or IV).
- Another scenario: 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, we still need to choose the cross-sectional weight instead of the longitudinal weight, the one where the outcome variable is from.

Cross-Sectional Weights
Single-Level (Population Average) Models (Individual-Level)

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 who were enrolled in US schools during the 1994-1995 academic year for the specified grades (Grade 7-12 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.	same as above
Wave III (2001)	GSWGT3_2 (N=14,322)	Wave I respondents who were interviewed at Wave III.	Same as above
Wave IV (2008)	GSWGT4_2 (N=14,800)	Wave I respondents who were interviewed at Wave IV.	Same as above

Choosing a Sampling Weight for Analysis
Longitudinal Analysis

- Longitudinal analysis is used to investigate research questions answered by investigating changes in measurements taken on subjects over time.
- The outcome variable is measured multiple times. *Note that if the covariates are from multiple waves but the outcome variable is just from one wave of data, this is NOT a longitudinal analysis*.
- 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 Weights
Single-Level (Population Average) Models – Individual Level

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 who were enrolled in US schools during the 1994-1995 academic year for the specified grades (Grade 7-12 in 1994-1995)
Wave IV (2008)	GSWGT4 (N=9,421)	Eligible Wave I respondents who were interviewed at Wave II, III & IV.	Same as above
Wave IV (2008)	GSWGT134 (N=12,288)	Eligible Wave I respondents who were interviewed at Wave III & IV.	Same as above

Note: this table only gives you some examples. See Table 2.5 in “Guidelines for Analyzing Add Health Data” for the complete list.

Choosing a Sampling Weight for Analysis
Time-to-Event (Survival) 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. Example events are death, onset of disease, first pregnancy, first marriage.
- The event may not be observed for all subjects. Choice of sampling weight will usually be determined by the data collected at the earliest time point.

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

	Data Source	Number in Analysis File	Weight for Individual-level Models
Data available from only one interview:			
Adolescents in 1995 enrolled in Grade7-12 during 1994-1995	Wave I only	18,924	GSWGT1
Adolescents in 1996 enrolled in Grade 7-11 during 1994-1995	Wave II only	13,570	GSWGT2
Young Adults in 2001 enrolled in Grade 7-12 during 1994-1995	Wave III only	14,322	GSWGT3_2
Young Adults in 2008 enrolled in Grade 7-12 during 1994-1995	Wave IV only	14,800	GSWGT4_2
Data available from Multiple interviews:			
Adolescents in 1995 enrolled in Grade7-12 during 1994-1995	Wave I & II	18,924	GSWGT1
Adolescents in 1996 enrolled in Grade 7-11 during 1994-1995	Wave II & III	13,570	GSWGT2
Young Adults in 2001 enrolled in Grade 7-12 during 1994-1995	Wave I, II, & III	14,322	GSWGT1
Young Adults in 2008 enrolled in Grade 7-12 during 1994-1995	Wave I, II, III & IV	14,800	GSWGT1

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.
- 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).
- Note that the two level sampling weights need to be scaled before you are running a multi-level model in different packages. Scaling methods may differ depending on what package you use.
- There are two different methods of scaling the sampling weights for estimating this model, PWIGLS METHOD 2 and MPML METHOD A.

Cross-Sectional Weight Components for Multi-Level Models
Cross-Sectional Analysis

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 I (1995)	SCHWT1 (N=132)	W1_WC (N=18,924)	Same as above	Same as above
Wave II (1996)	SCHWT1 (N=132)	W2_WC (N=13,568)	Same as above	Same as above
Wave III (2001)	SCHWT1 (N=132)	W3_2_WC (N=14,322)	Same as above	Same as above
Wave IV (2008)	SCHWT1 (N=132)	W4_2_WC (N=14,800)	Same as above	Same as above

Longitudinal Weight Components for Multi-Level Models Longitudinal Analysis

Data Used	Level 2 Weight Component (N)	Level 1 Weight Component (N)
WII & III	SCHWT1 (N=132)	W3_WC (N=10,828)
W I, II, & III	SCHWT1 (N=132)	W3_WC (N=10,828)
Wave I, II, III, & IV	SCHWT1 (N=132)	W4_WC (N=9,421)

See Table 2.5 for complete list in the "Guidelines for Analyzing Add Health Data"

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

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

Sampling Weights for Wave I Genetic Sample*
Estimating Single-Level (population average) Models

Data Set (Year collected)	Sampling Weight Variable (N)	Sample	Target Population
Wave I (1995)	PERSONWEIGHT (N=5,530)	Genetic sample of individuals with varying genetic resemblance, including monozygotic twins, dizygotic twins, full siblings, half siblings, and unrelated siblings who were raised in the same household.	1995 US population of persons age 12 to 18 who live under the same household.
	PAIRWEIGHT (N=3,160)	Genetic sample of pairs with varying genetic resemblance, including monozygotic twins, dizygotic twins, full siblings, half siblings, and unrelated siblings who were raised in the same household.	1995 US population of pairs of individuals age 12 to 18 who live under the same household.
Note: *Users do not need to use genetic sample weights if you use data from this supplemental sample.* Anyone who is interested in using weights for the genetic sample can contact the Add Health support group to request the weights and corresponding documentation. Anyone who wants to use the weights needs to have a good understanding of and agrees with the weighting procedure.			

Choosing a Sampling Weight for Analysis
Common Errors to Avoid

- Do *NOT* normalize the weights (by dividing the survey weight of each unit by the (unweighted) average of the survey weights of all the analyzed units) 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.

- Do not use frequency or analytical weight (in Stata fweight(); aweight())

  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.

Choosing a Sampling Weight for Analysis
Common Errors to Avoid

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. The analyst should be sure the Add Health weights are used as sampling weights. (For example, in Stata, use pweight() option.)

Software packages do not always give different statements that uniquely define the type of weight. For example, the SAS statement:

WEIGHT GSWGT1 - a frequency weight in PROC FREQ

a variance weight in PROC REG

a sampling weight in PROC SURVEYREG

Things to Cover

- Special features of Add Health design
- How to choose the correct sampling weight for analysis
- Steps of Preparing Data for Analysis
- Examples
  - Descriptive Statistics
  - OLS regression
  - Subpopulation Analysis
  - Multilevel Model

Steps of Preparing Data for Analysis

- Determine the wave(s) of data you need for your analysis and construct desired variables.
- Identify the attributes & elements of the sample design (with replacement Design, strata variable, cluster variable, weight variable) for the data identified in step 1.

Design Type: Specify With Replacement as the Design Type

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 with replacement sampling.

Stratum Variable: Use REGION

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. 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.

Steps of Preparing Data for Analysis

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

This is the variable named PSUSCID for the In-School, Wave I, II, III, and IV 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 do and choose an appropriate weight variable according to the guidelines provided in previous slides.

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

Therefore, we not only need to include the weight variable in our analysis, but also need to include cluster and stratification variables to correct for design effects; otherwise, variance estimates may not be correct.

- Delete cases that have a missing value for a sampling weight. Otherwise the sample size is not correct. However, you cannot delete cases that do not belong to your subsample (if statement = listwise deletion). You need to use subpopulation analysis.

Things to Cover

- Special features of Add Health design
- How to choose the correct sampling weight for analysis
- Steps of Preparing Data for Analysis
- Examples
  - Descriptive Statistics
  - OLS regression
  - Subpopulation Analysis
  - Multilevel Model

Example 1. Example for Descriptive Statistics.

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

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

SAS 9.2.3 syntax:

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

STATA 12.1 syntax:

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

UNC CAROLINA POPULATION CENTER

Add Health
The National Longitudinal Study of Adolescent to Adult Health

Example 1. Example for Descriptive Statistics.

Table 4.1 Parameter estimates and standard errors to predict the average number of hours TV watched during a week by adolescents.

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

Things to Cover

- Special features of Add Health design
- How to choose the correct sampling weight for analysis
- Steps of Preparing Data for Analysis
- Examples
  - Descriptive Statistics
  - OLS regression
  - Subpopulation Analysis
  - Multilevel Model

Example 2. Regression Example for Population-Average Models

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)?

STATA 12.1 syntax:

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

SAS 9.1 syntax:

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

UNC CAROLINA POPULATION CENTER

Add Health
The National Longitudinal Study of Adolescent to Adult Health

Example 2. Regression Example for Population-Average Models

Table 4.3 Parameter estimates and standard errors to predict the percentile score on the Add Health PVT test.

Parameter	SAS 9.2.3 Estimate (Std Err)	Stata 12.1 Estimate (Std Err)
β_0 (INTERCEPT)	69.946 (7.855)	69.946 (7.854)
β_1 (AGE_W1)	-1.085 (0.489)	-1.085 (0.489)
β_2 (BOY)	3.395 (0.673)	3.395 (0.673)
β_3 (HR_WATCH)	-0.150 (0.020)	-0.150 (0.020)

Things to Cover

- Special features of Add Health design
- How to choose the correct sampling weight for analysis
- Steps of Preparing Data for Analysis
- Examples
  - Descriptive Statistics
  - OLS regression
  - Subpopulation Analysis
  - Multilevel Model

Subpopulation Analysis

- Subsample (e.g. female; African American; Asian); missing data; merging multiple data sets

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

- Common errors – DELETING CASES:
  - Delete (drop) cases that do not belong to your analysis sample
  - Doing nothing; meaning listwise deletion
  - Use “if” statement to specify the subsample

Subpopulation Analysis

Data Set (Year collected)	Sampling Weight Variable (N)	N
Wave I (1995)	GSWGT1	18,924
Wave II (1996)	GSWGT2	13,570
Wave III (2001)	GSWGT3_2	14,322
Wave IV (2008)	GSWGT4_2	14,800

Subpopulation Analysis

Data Set (Year collected)	Sampling Weight Variable (N)	N
Wave I, II, III	GSWGT3	10,828
Wave I, II, III, IV	GSWGT4	9,421
Wave I, III, IV	GSWGT134	12,288

Subpopulation Analysis

- If observations are deleted from the data set, the standard errors of the estimates might 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 deleting cases that are not in your analysis sample, 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(s) 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

- The size of difference in the two variance estimates from analyzing the full dataset with the subpopulation option and the subset of the data is hard to predict. If only a few PSUs are missing in each level of the stratification variable (REGION), then your results will probably be nearly the same.
- Make sure that all PSUs are represented in each level of the stratification variable.
- 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, so that 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 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

- Before you do the analysis, you often need to prepare a subpopulation variable. Suppose you are interested in studying a subgroup 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 subpopulation option in your analysis.

Subpopulation Analysis

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

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

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

STATA 12.1 CORRECT way of using SUBPOP option

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

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

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

SAS 9.2.3 syntax for using DOMAIN statement to specify subpopulation

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

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

Subpopulation Analysis

- SAS does allow users to specify subpopulations with the DOMAIN statement in PROC SURVEYMEANS.
- Stata has a subpopulation option for use.
- However, none of the other SAS SURVEY procedures allow users to analyze subpopulations. But the SAS SURVEY software can be tricked into computing the correct variance and standard errors when analyzing subpopulations. If you are using a newer version of SAS (like 9.3), surveyreg then has the “domain” option.
- ***Set weights outside the subpopulation to a very small value (close to zero). Note*** SAS delete observations that have a zero value for the sampling weight. So do not use zero value for weights.

Subpopulation Analysis – Example 2. Multivariate Analysis

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

STATA 12.1 with correct subpopulation option

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

SAS 9.2.3 version 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;
```

Subpopulation Analysis – Example 2. Multivariate Analysis

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

SAS 9.2.3 version 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)

Things to Cover

- Special features of Add Health design
- How to choose the correct sampling weight for analysis
- Steps of Preparing Data for Analysis
- Examples
  - Descriptive Statistics
  - OLS regression
  - Subpopulation Analysis
  - Multilevel Model

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).

Choosing a Sampling Weight for Analysis
Multilevel Model

- In a single (usually) level model, we only need to use grand sample weight (wij), because the grand sample weight factors in all levels of clustered sampling, corrections for nonresponse, oversampling, and post-stratification. Wij is an unconditional weight for observation i,j.
- In a two-level model, it is not it is not sufficient to use the single grand sampling weight w_{ij}, because weights enter into the log likelihood at both the school level and individual level.
- Instead, what is required for a two-level model under this sampling design is w_j (the inverse of the probability that school *j* is selected in the first stage), and $w_{i|j}$ (the inverse of the probability that individual *i* from school *j* is selected at the second stage conditional on school *j* already being selected). It is not appropriate to use grand sample weight w_{ij} without making assumption about w_j.
- Both the school-level w_j and individual-level $w_{i|j}$ are called weight components in Add Health. As mentioned earlier, if both the school-level and individual-level weight components are included in the two-level model, rescaling is necessary to remove the dependence of $w_{i|j}$ on w_j.
- Further details on weighting and scaling in xtmixed with Survey data are available in the Stata manual (p. 342-343).

Scaling Sampling Weights

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

$$mp\_wt\_w1_{i,j} = \frac{w1\_wc_{i|j} * schwt1_j}{\left(\frac{\sum_{i}^{n_1} w1\_wc_{i|j}}{n_j} \right)}$$

- This is just the product of the PWIGLS scaled level 1 weight and the level 2 weight. The analyst 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 Multi-level Model

	Use PWIGLS Method 2	Need to use PWIGLS program to do the scaling before running the multi-level model	Use MPML Method A	Need to use MPML_WT program to do the scaling before running the multi-level model
XTMIXED in Stata	Yes	No. Instead, use "pwscale(size)" option in XTMIXED	No	NA
GLLAMM in Stata	Yes	Yes	No	NA
LISREL	Yes	No	No	NA
MLWIN	Yes	No	No	NA
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 help doing the needed scaling of the weights. See appendix A in Guidelines.**

Scaling Weights for Multilevel Analysis

Appendix A (p. 46-47)

(From Chen and Chantala. 2014 “Guidelines for Analyzing Add Health Data.”

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

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

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

Also available from the CPC website (http://www.cpc.unc.edu/research/tools/data_analysis) is documentation that provides

(1) information on using these programs to create the two-level weights

(2) information about several popular multilevel software packages that allow these sampling weights to be used in estimation

(3) instructs the analyst in downloading and running these programs.

Example Code Used to Construct Weights for gllamm Cross-Sectional Analysis

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

- 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 Cross-Sectional Analysis

- 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.

Example Code Used to Construct Weights for Mplus Cross-Sectional Analysis

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)
```

UNC CAROLINA POPULATION CENTER

Add Health
The National Longitudinal Study of Adolescent to Adult Health

Scaling Weights for Multilevel 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 option mpml_wta will generate the weight variable “mp_wt_w1” for use in estimating 2-level models in Mplus.

Example for Multi-Level Model

- 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.
- 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).

Example for Multi-Level Model

- School-level independent variable: the availability of an on-site school recreation center (variable RC_S).

Student-level model (Within or Level 1):

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

where:

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

School-level Model (Between or Level 2):

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

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

where:

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

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;
```

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

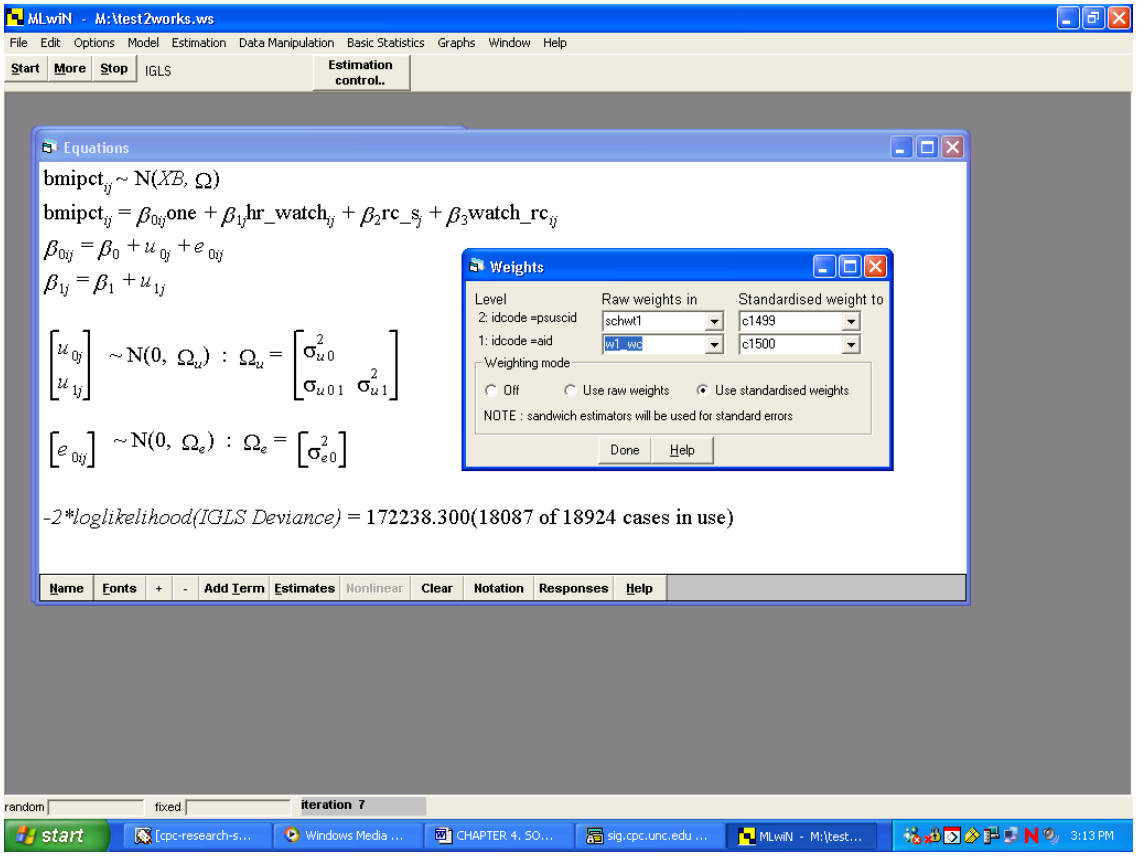
LISREL

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

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

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

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



UNC CAROLINA POPULATION CENTER

Add Health
The National Longitudinal Study of Adolescent to Adult Health

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 ($\delta_{0j}\delta_{1j}$))	-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)

Program Syntax for Multilevel Analysis – Similar Data Set

XTMIXED (in Stata 12.1)

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

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

Program Syntax for Multilevel Analysis – Similar Data Set

```
MPLUS 4.0

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

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

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

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)

Methodology Session: Modeling Contextual Data in Add Health

Friday, June 27, 2014 10:15 Breakout Session 5

Presenter: Sharon Christ, Purdue University

Abstract

This presentation provides an overview of two approaches to modeling Add Health data utilizing the contextual variables available in the Wave I, II, III, and IV Contextual files and the Wave I and III ONE files. Multilevel and marginal modeling approaches will be discussed including longitudinal modeling within these frameworks. Special focus will be on how to properly adjust model estimates for the complex sample design of the Add Health, including applying the correct sampling weights and correcting for the non-independence (clustering) of observations. Treatment of missing data will also be touched upon due to its relationship to sample weight selection. Pros and Cons of the different modeling approaches and estimation methods will be considered.

- ✓ Neighborhood level data (state-level; county-level; tract-level; block-level)
- ✓ Longitudinal modeling; Growth curve model (time nested within individuals)
- ✓ Three level analysis: time nested within individuals nested within neighborhoods

Summary - Things to Remember

- Add Health is a national longitudinal study that has special survey design features.
- Users need to account for those special features, including clustering, stratification, and unequal probability of selection, when they analyze Add Health data. Otherwise, point estimates and variances may be biased; and inferences drawn from the results are not correct.
- Use cluster variable (psuscid), stratification variable (region), and weight variables.
- Choose the correct sampling weight variable for different types of analysis.
- Use subpopulation option.
- Use two-level weight components variables and scale them when analyzing a school-level and individual-level model.

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