Modeling Contextual Data in the Add Health

Sharon L. Christ
Departments of HDFS and Statistics
Purdue University

UNC
CAROLINA
POPULATION
CENTER

Talk Outline

1. Review of Add Health Sample Design

2. Modeling Add Health Data
   a. Multilevel and Marginal Modeling
   b. Longitudinal Modeling

3. Contextual Data available in Add Health
4. Incorporating Contextual Data into Models

Sample Design Review

Add Health is a school-based design.

80 High schools selected with probabilities proportionate to size (PPS) where size is the number of students

Feeder (middle) schools were selected PPS for each high school

132 schools are in the sample (some schools spanned grades 7 to 12)

School sizes vary from 100 students to 3,000 students

Sample Design Review

Schools were selected with unequal probability.

A PPS design means that larger schools had a higher probability of being selected into the sample of schools.

Therefore, the sample of schools is not representative of the population of schools.

Corrections are needed to obtain a representative sample of schools. (sample weights)

Sample Design Review

PPS example

If there are two schools in the population and school 1 had 1,000 students and school 2 had 500 students, then the probability of selection for each school is:

school 1 = 1,000/1,500 = 1/1.5

school 2 = 500/1,500 = 1/3

Sample Design Review

Students were selected from the sampled schools.

Approximately 200 students from each school pair were selected for the core sample (n=12,105).

Additional supplemental samples of adolescents were selected based on specific criteria.

The total sample (core + supplemental) is made up of 20,745 adolescents.

Sample Design Review

Adolescents within the same school are not independent

Selection of an adolescent into the sample depends on their school having been selected.

Adolescents within the same school are not independent of one another.

Adolescent outcomes will be clustered, more similar, within schools than across schools.

Sample Design Review

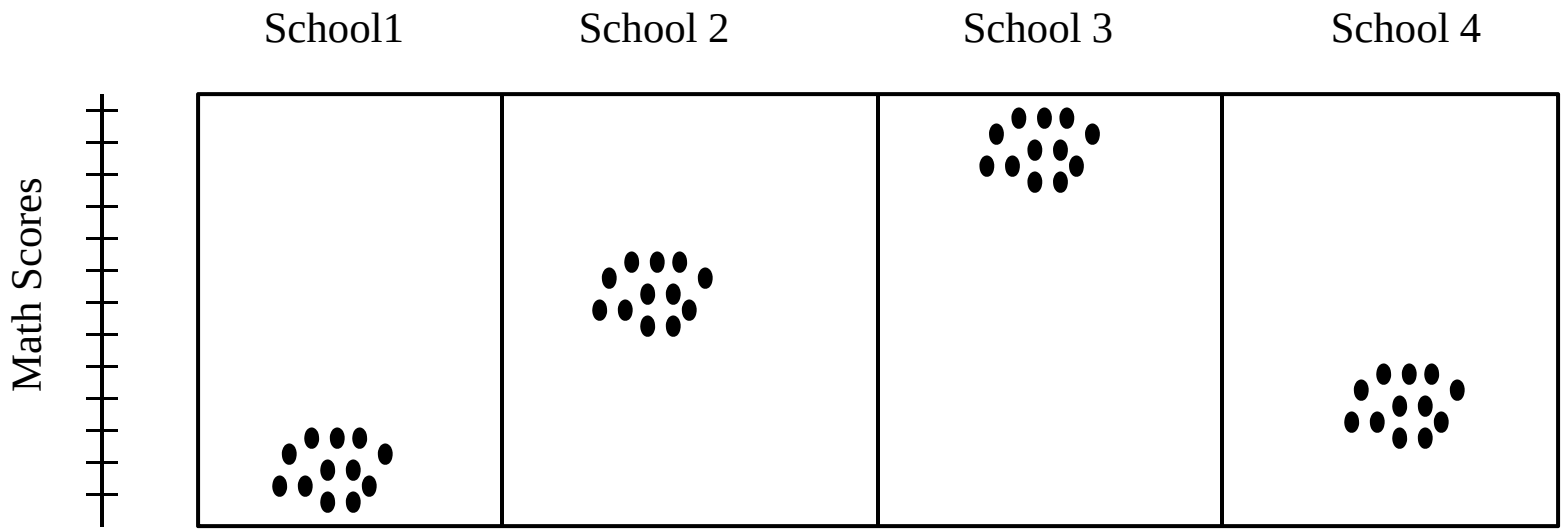
Adolescents within the same school are not independent

Clustering occurs to different degrees depending on the outcome under study.

Clustering also likely weakens at later waves during young adulthood when schools are less salient.

Sample Design Review

High Degree of Clustering Example:

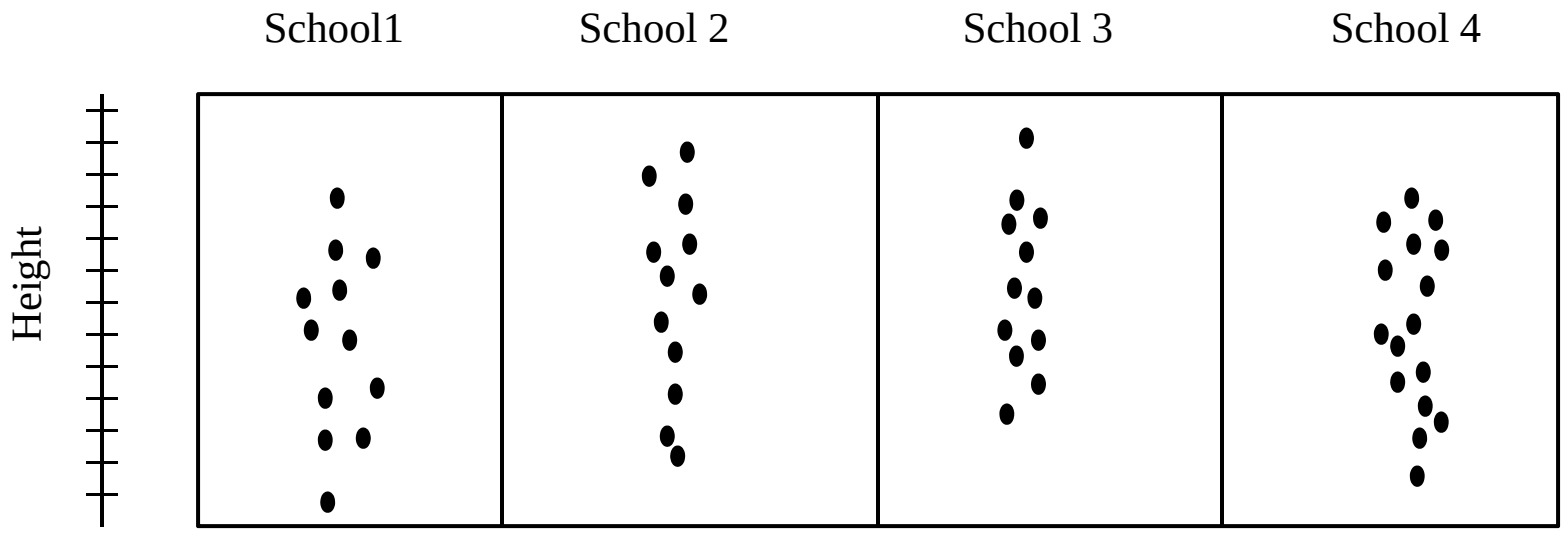


UNC CAROLINA POPULATION CENTER

Add Health
The National Longitudinal Study of Adolescent to Adult Health

Sample Design Review

Low Degree of Clustering Example:



UNC CAROLINA POPULATION CENTER

Add Health
The National Longitudinal Study of Adolescent to Adult Health

Sample Design Review

The intra-class correlation coefficient estimates the degree of clustering in the population.

$$\text{ICC} = \hat{\rho} = \text{between school variance} \Big/ \text{total variance}$$

Always check this for your outcome to understand the proportion of variance that may be explained by school level factors.

Sample Design Review

Example ICC's in Add Health (weighted)

parent education	0.186
school is safe	0.109
self-rated health	0.027
mental health (depression)	0.022

Sample Design Review

Adolescents from the combined (core + supplemental) sample are selected with unequal probability

Approximately equal number of students from each school were randomly selected into the core sample.

Because schools differ in size, students from larger schools have a smaller probability of being in the sample.

In addition, stratification was used to over-select students meeting the special criteria for the supplemental files.

Sample Design Review: Corrections

Add Health sampling weights are designed to correct bias resulting from unequal probabilities of selection

Sampling weights are the inverse of the probability of selection of an observation.

Example: Probability of selection into a sample is 200 out of 1,500 or 0.13 for observation *i* in school 1, the weight for observation *i* is

$$\frac{1}{0.13} = 7.7$$

Sample Design Review: Corrections

In Add Health, the sample is nearly self-weighting for the core sample.

Example:

School weight	Student weight	Final weight
1,500/1,000	1,000/200	1,500/200
1,500/500	500/200	1,500/200

Sample Design Review: Corrections

Sampling weights correct for biases, but add imprecision (increase variance in) estimates

The unequal weighting effect (UWE), is the estimated increase in the standard error of a mean due to the variance in the weights.

$$UWE = 1 + \frac{Var(w_i)}{\bar{w}^2} = 1 + cv^2_{weights}$$

where w_i is the weight for individual *i*

$cv^2_{weights}$ is the squared coefficient of variation of the weights

UNC CAROLINA POPULATION CENTER

Sample Design Review: Corrections

UWE for Add Health Cross Sectional and Panel Weights

The square-root of the UWE for cross sectional weights ranges between 1.32 and 1.37 indicating that the standard error for a mean estimate is about 1.3 times larger than it would have been in a equal probability sample, i.e., not using weights.

Similar, but slightly higher values (1.36 – 1.39) for the panel (longitudinal) weights.

Sample Design Review: Corrections

Clustering of students within schools violates common analytic assumption of independent observations.

Clustering biases standard errors. Two approaches to correcting this bias are common in statistical analysis:

1) Marginal or “Population Average” modeling, which treats clustering as a nuisance and uses standard error estimators that are robust to clustering.

2) Multilevel (Hierarchical, Mixed-Effects, Random-Effects) modeling, which explicitly estimates the variance components at the cluster and within-cluster levels.

Longitudinal Nature of Add Health

Wave I 1994-1995	Wave II 1996	Wave III 2001-2002	Wave IV 2008
Adolescents grades 7-12 n = 20,745	Adolescents grades 8-12 n = 14,738	Young Adults aged 18-26 n = 15,197	Adults Aged 24-32 n = 15,701

UNC
CAROLINA
POPULATION
CENTER

Add Health
The National Longitudinal Study of Adolescent to Adult Health

Longitudinal Nature of Add Health

Longitudinal weights further adjust for unit (person) attrition

Types of weights:

1) Cross-sectional weights for use when analyzing individuals from a single wave of data
2) Panel (longitudinal) weights used when analyzing individuals in 2 or more waves of data

Longitudinal Nature of Add Health

Multi-level weights

Types of weights:

3) Multi-level cross-sectional weights for use when analyzing a single wave of data in an MLM
4) Multi-level panel (longitudinal) weights used when analyzing individuals in 2 or more waves of data in an MLM

Longitudinal Models

UWE for Add Health Multilevel Weights

The square-root of the UWE for the school weight is 1.6. Indicating that the standard error for a mean estimate is about 1.6 times larger due to unequal selection of schools.

The square-root of the UWE for level-1 (conditional student) weights ranges between 1.54 - 1.6.

Therefore, less efficient than single-level weights.

Longitudinal Models

Panel and Change Models

A **panel model** uses variables from more than one wave of data. For example, predicting outcomes at wave IV using predictors from waves I and II. Outcomes are generally from one wave of data.

A **longitudinal (change) model** also uses variables from more than one wave, but observations are person-by-time where outcomes from multiple waves are modeled over time.

UNC CAROLINA POPULATION CENTER

Data Structures for Longitudinal Models

observation (“wide”) data structure – used for panel models

School ID	Respondent ID	depression WI	depression WII	depression WIII	depression WIV
1	1	12	11	12	13
1	2	10	10	9	8
2	3	7	.	8	.
2	4	8	6	7	6
2	5	14	10	11	.

School ID	Respondent ID	Wave	Depression
1	1	1	12
1	1	2	11
1	1	3	12
1	1	4	13
1	2	1	10
1	2	2	10
1	2	3	9
1	2	4	8
2	3	1	7
2	3	3	8
2	4	1	8
2	4	2	6
2	4	3	7
2	4	4	6
2	5	1	14
2	5	2	10
2	5	3	11

observation-by-time ("stacked" or "long") data structure

generally used for longitudinal (change) models.

Longitudinal Models

Data structure and the treatment of missing data affects which weight you should use.

If you do a case-wise deletion with a wide data structure, use the panel weights (Add Health longitudinal weights) appropriate for the waves you are including.

If you will use maximum-likelihood (a.k.a., FIML) or multiple imputation (MI), with a wide data structure, then you may use a cross sectional weight in your model.

UNC CAROLINA POPULATION CENTER

Add Health
The National Longitudinal Study of Adolescent to Adult Health

Longitudinal Models

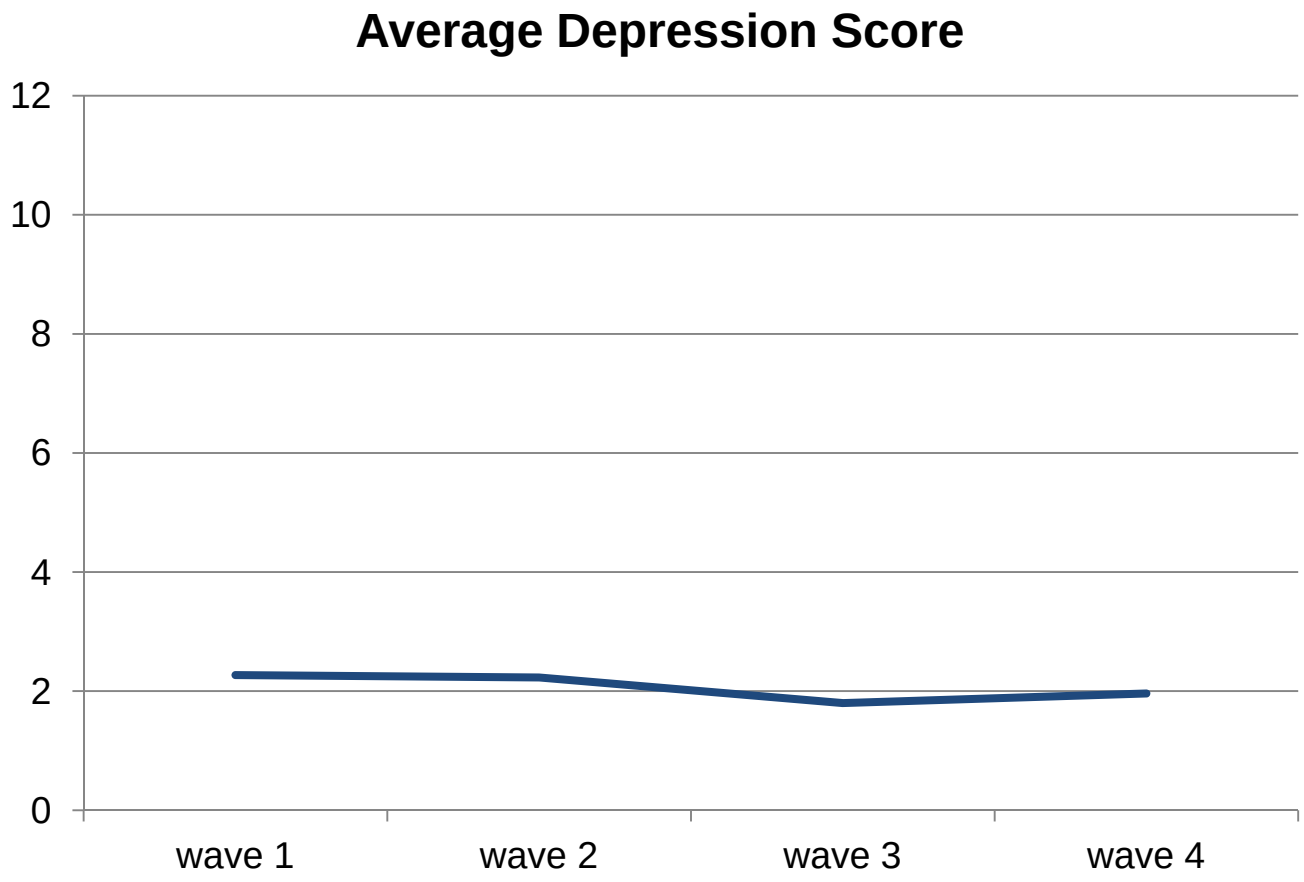
Data structure and the treatment of missing data affects which weight you should use.

In a stacked data structure where you are evaluating change and persons are missing at some waves and not others, use a time-varying weight.

Time-varying weights are the cross sectional weights for specific waves.

School ID	Respondent ID	Wave	Time-varying weight	Depression
1	1	1	wave 1 weight	12
1	1	2	wave 2 weight	11
1	1	3	wave 3 weight	12
1	1	4	wave 4 weight	13
1	2	1	wave 1 weight	10
1	2	2	wave 2 weight	10
1	2	3	wave 3 weight	9
1	2	4	wave 4 weight	8
2	3	1	wave 1 weight	7
2	3	3	wave 3 weight	8
2	4	1	wave 1 weight	8
2	4	2	wave 2 weight	6
2	4	3	wave 3 weight	7
2	4	4	wave 4 weight	6
2	5	1	wave 1 weight	14
2	5	2	wave 2 weight	10
2	5	3	wave 3 weight	11

Time-Varying Weights



UNC CAROLINA POPULATION CENTER

Add Health
The National Longitudinal Study of Adolescent to Adult Health

Talk Outline

1. Review of Add Health Sample Design

2. Modeling Add Health Data
   a. Multilevel and Marginal Modeling
   b. Longitudinal Modeling

3. Contextual Data available in Add Health
4. Incorporating Contextual Data into Models

Two Types of Models

Two theoretical approaches to modeling data from complex samples

Design-based Approach: The complex sample is a nuisance. Use estimators that are robust to nesting. Use weights to correct for unequal selection probabilities.

Model-based Approach: The model should be correct and if correct, it is robust to the complex sample design. Variables related to selection and clustering are included explicitly in the model.

UNC CAROLINA POPULATION CENTER

Two Types of Models

The nesting of students within schools may be dealt with implicitly using **marginal or "population average"** modeling

Generalized Estimation Equation (GEE) is used to refer to a marginal model as well.

These models are in the design-based tradition and often used in demographic and public health disciplines.

Sampling statisticians generally design samples, like the Add Health sample, with these types of models in mind.

UNC CAROLINA POPULATION CENTER

Two Types of Models

Marginal models are single-level models.

Clustering or nesting does not bias point estimates, but does bias standard error estimates and test statistics.

Marginal models simply use a variance (standard error) estimator that is robust to the nesting to correct for this problem.

Use single-level weights with this model.

Two Types of Models

Marginal models do not explicitly model error terms.

They model the average effects for the population:

$$E\left(y_{ij}\right) = \alpha + \sum \beta_p x_{pij}$$

e.g.,

$$E\left(\text{depression}_{ij}\right) = \alpha + \beta_1 \text{age}_{ij} + \beta_2 \text{sex}_{ij} + \beta_3 \text{close}_{ij}$$

where *E* is the expected value.

Two Types of Models

$$E\left(\text{depression}_{ij}\right) = \alpha + \beta_1 \text{age}_{ij} + \beta_2 \text{sex}_{ij} + \beta_3 \text{close}_{ij}$$

mentalw1	Coef.	Linearized Std. Err.	t	P>|t|	[95% Conf.	Interval]
ageyw1	.1108823	.0171734	6.46	0.000	.0769093	.1448553
male	-.7178441	.0512822	-14.00	0.000	-.8192925	-.6163958
closew1	-.6210313	.0380343	-16.33	0.000	-.6962722	-.5457904
_cons	3.714684	.3569225	10.41	0.000	3.008606	4.420762

svy commands in STATA

proc survey commands in SAS

UNC CAROLINA POPULATION CENTER

Two Types of Models

The nesting of students within schools may be dealt with explicitly using a **multilevel** (hierarchical, mixed-effects, random-effects) model.

These models are in the model-based tradition and used in economics, psychology, and other social science disciplines.

Sampling statisticians do not generally design samples with these types of models in mind. However, Add Health created weights for these models post-hoc.

UNC CAROLINA POPULATION CENTER

Add Health
The National Longitudinal Study of Adolescent to Adult Health

Two Types of Models

Multi-level models have, you guessed it, multiple levels.

Multilevel models explicitly estimate the variance at each level (school and student).

Standard errors are adjusted for the clustering.

Typically in these models you want to be able to make predictions for a specific individual rather than just estimate means for the population.

Also, you may want to evaluate variance components.

UNC CAROLINA POPULATION CENTER

Two Types of Models

Multilevel models do explicitly model error terms, disaggregated across levels.

They model the average effects for the population AND the heterogeneity around those averages:

$$y_{ij} = \alpha + \sum \beta_p x_{pij} + \mu_{0j} + \varepsilon_{ij}$$

e.g.,

$$\text{depression}_{ij} = \alpha + \beta_1 \text{age}_{ij} + \beta_2 \text{sex}_{ij} + \beta_3 \text{close}_{ij}$$
$$+ \text{resvar}(\text{school}) + \text{resvar}(\text{student})$$

Two Types of Models

mentalw1	Coef.	Robust Std. Err.	z	P>|z|	[95% Conf.	Interval]
ageyw1	.11239	.0174344	6.45	0.000	.0782191	.1465608
male	-.6838686	.0603143	-11.34	0.000	-.8020823	-.5656548
closew1	-.6155772	.0518039	-11.88	0.000	-.7171109	-.5140435
_cons	3.576686	.4051791	8.83	0.000	2.78255	4.370823

Random-effects Parameters	Estimate	Robust Std. Err.	[95% Conf.	Interval]
PSUSCID: Identity				
var(_cons)	.0906775	.0225791	.0556605	.1477245
var(Residual)	4.629412	.1353261	4.371632	4.902391

xtmixed, xtgee commands in STATA

proc mixed commands in SAS

UNC CAROLINA POPULATION CENTER

Add Health

The National Longitudinal Study of Adolescent to Adult Health

Two Types of Models

Compare results for the “fixed effects”:

	Marginal Model Results		Multilevel Model Results	
	beta	z	beta	z
age	0.111	6.5	0.112	6.5
male	-0.718	-14.0	-0.684	-11.3
close	-0.621	-16.3	-0.616	-11.9
intercept	3.715	10.4	3.577	8.8

More efficiency with marginal model? – weights?

UNC CAROLINA POPULATION CENTER

Add Health
The National Longitudinal Study of Adolescent to Adult Health

Two Types of Models: Longitudinal

Both types of models may be extended to longitudinal cases

For “panel models” that involve an outcome from one time point, models are the same as shown above.

For “longitudinal models” that model an outcome over time from multiple waves (i.e., repeated measures), the marginal model remains a single-level model and the multi-level model becomes a three-level model.

School ID	Respondent ID	Wave	Depression
1	1	1	12
1	1	2	11
1	1	3	12
1	1	4	13
1	2	1	10
1	2	2	10
1	2	3	9
1	2	4	8
2	3	1	7
2	3	3	8
2	4	1	8
2	4	2	6
2	4	3	7
2	4	4	6
2	5	1	14
2	5	2	10
2	5	3	11

Two levels of nesting:

Students nested within schools

Time nested within students

UNC CAROLINA POPULATION CENTER

Add Health
The National Longitudinal Study of Adolescent to Adult Health

Two Types of Models: Longitudinal

Marginal model of change in depression over time

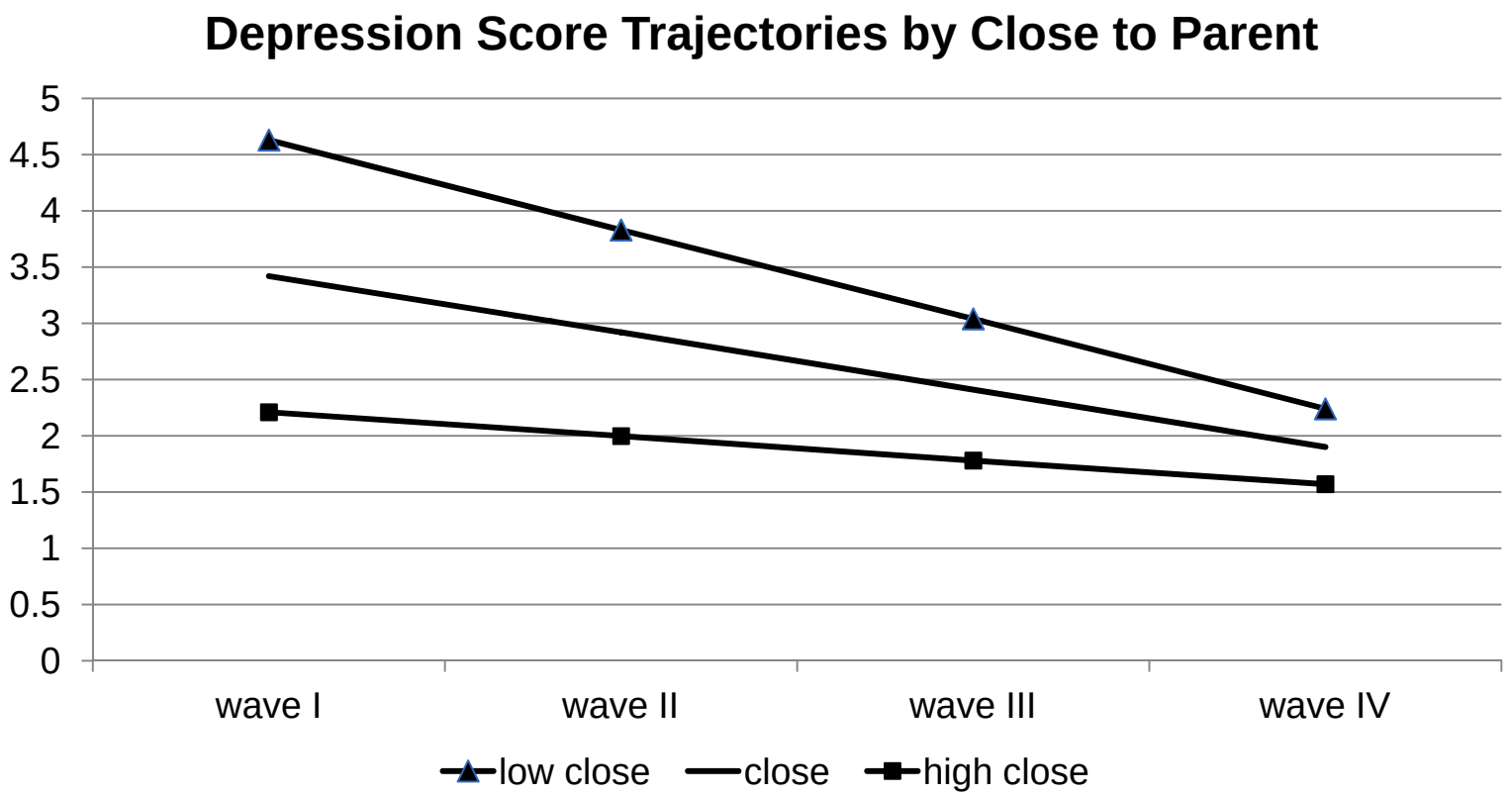
$$E\left(\text{depression}_{tij}\right) = \alpha + \beta_1 \text{wave}_{tij} + \beta_2 \text{age}_{tij} + \beta_3 \text{sex}_{ij} + \beta_4 \text{close}_{ij} + \beta_5 \text{close}_{ij} * \text{wave}_{tij}$$

mental	Coef.	Linearized Std. Err.	t	P>|t|	[95% Conf.	Interval]
wave	-.9440002	.1944157	-4.86	0.000	-1.328714	-.5592866
age	.0372978	.0148352	2.51	0.013	.0079416	.066654
male	-.4580876	.0829547	-5.52	0.000	-.62224	-.2939353
closew1	-.6046271	.0733058	-8.25	0.000	-.749686	-.4595682
inter	.1456013	.0414911	3.51	0.001	.063498	.2277047
_cons	4.692899	.4065576	11.54	0.000	3.888395	5.497403

UNC CAROLINA POPULATION CENTER

Add Health
The National Longitudinal Study of Adolescent to Adult Health

Two Types of Models: Longitudinal



UNC CAROLINA POPULATION CENTER

Add Health
The National Longitudinal Study of Adolescent to Adult Health

Two Types of Models: Longitudinal

Multi-level model of change in depression over time

$$\text{depression}_{tij} = \alpha + \beta_1 \text{wave}_{tij} + \beta_2 \text{age}_{tij} + \beta_3 \text{sex}_{ij} + \beta_4 \text{close}_{ij} + \beta_5 \text{close}_{ij} * \text{wave}_{tij}$$
$$+ \text{resvar}(\text{school}) + \text{resvar}(\text{student}) + \text{resvar}(\text{wave})$$

mental	Coef.	Robust Std. Err.	z	P>|z|	[95% Conf.	Interval]
wave	-1.010767	.1742884	-5.80	0.000	-1.352366	-.6691683
age	.0226453	.0147138	1.54	0.124	-.0061932	.0514838
male	-.4214568	.0890596	-4.73	0.000	-.5960105	-.2469031
closew1	-.6596564	.0800711	-8.24	0.000	-.816593	-.5027199
inter	.1737519	.0342774	5.07	0.000	.1065694	.2409344
_cons	5.138527	.435925	11.79	0.000	4.28413	5.992924

UNC CAROLINA POPULATION CENTER

Add Health
The National Longitudinal Study of Adolescent to Adult Health

Two Types of Models: Longitudinal

Multi-level model of change in depression over time

Random-effects Parameters	Estimate	Robust Std. Err.	[95% Conf.	Interval]
AID: Identity				
var(_cons)	2.699907	.1644899	2.396018	3.04234
PSUSCID: Identity				
var(_cons)	7.51e-25	9.21e-24	2.75e-35	2.05e-14
var(Residual)	2.25694	.0741777	2.116138	2.407111

*note, these estimates likely biased

UNC CAROLINA POPULATION CENTER

Add Health
The National Longitudinal Study of Adolescent to Adult Health

Two Types of Models: Longitudinal

Compare results for the “fixed effects”:

	Marginal Model Results		Multilevel Model Results	
	beta	z	beta	z
wave	-0.944	-4.9	-1.010	-5.8
age	0.037	2.5	0.023	1.5
male	-0.458	-5.5	-0.422	-4.7
close	-0.605	-8.3	-0.660	-8.2
close*wave	0.146	3.5	0.174	5.1
intercept	4.693	11.5	5.138	11.8

Two Types of Models: Longitudinal

Recommendation: use the marginal modeling approach for longitudinal analyses

Add Health does not provide 3-level weights and STATA would not scale weights for a 3-level model.

I used school weights at level 3 (school) and conditional respondent weights at level 2 (person) for the example.

MLM took much longer to estimate and is computationally intensive.

Talk Outline

1. Review of Add Health Sample Design

2. Modeling Add Health Data
   a. Multilevel and Marginal Modeling
   b. Longitudinal Modeling

3. Contextual Data available in Add Health
4. Incorporating Contextual Data into Models

Contextual Data

Schools are an obvious context of interest in analysis of adolescent development.

Add Health contains variables about the school measured at both the school, e.g., whether it is public or private, and student levels, e.g., student's views on the safety of their school.

These school variables may be included in a marginal or a multi-level model.

Contextual Data

The advantage in the multilevel model is that one may evaluate heterogeneity at the student and school levels and PRE type effect sizes.

However, you may still evaluate “individual differences” in terms of moderated effects within a marginal model context.

Most of our hypotheses are about the fixed effects anyway.

Contextual Data

Clustering within schools is likely due in part to the school and in part to the school neighborhood. Schools are a proxy for a geographic or spatial area.

Add Health has included many interesting contextual variables at different geographic levels: state, county, census tract, and census block group levels.

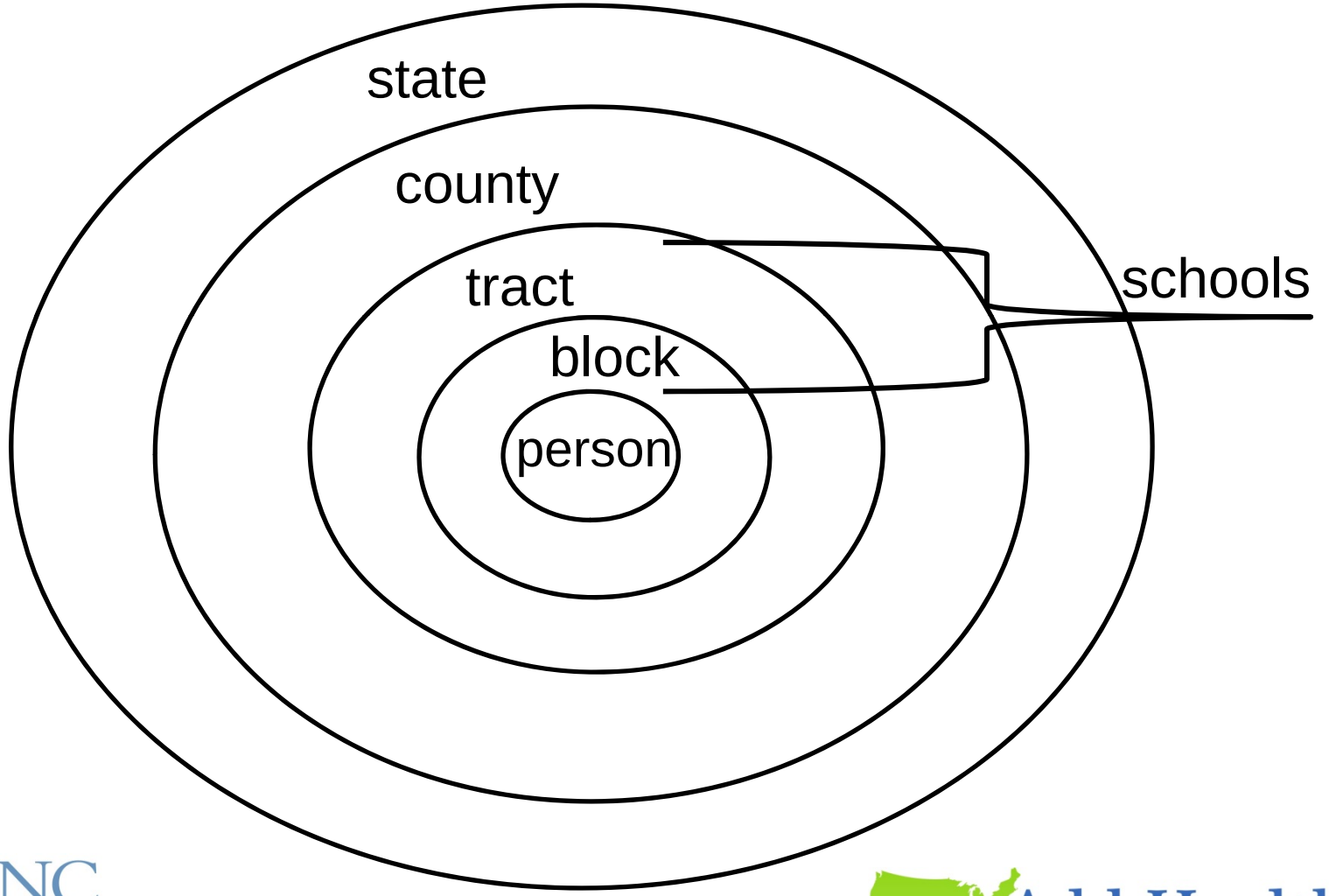
There is no violation of the independence assumptions at these other geographic levels, only for schools.

Contextual Data

A block group is a “subdivision of a census tract… [and] consists of all the blocks within a census tract with the same beginning number.”1 Block groups average about 1000 inhabitants.

A census tract is a “small, relatively permanent statistical subdivision of a county ….designed to be relatively homogenous units with respect to population characteristics, economic status, and living conditions at the time of establishment, census tracts average about 4,000 inhabitants.

Contextual Data



Contextual Data

Compare ICCs across contexts (unweighted)

	depression	SRH
school	0.021	0.017
block	0.026	0.032
tract	0.021	0.024
county	0.016	0.017
state	0.006	0.007

Contextual Data

Waves I, II, and III

Contextual variables including rates, proportions, and population measures are linked to respondent ID's based on their address.

1. Most of the variables are county level at Waves I and II. Wave III has more tract and block group variables.
2. Wave III has many of the same variables, but not all.
3. Therefore, some variables are time varying across Waves I, II, and III.

Contextual Data

Example Variables (of hundreds):

Child/woman ratio

Arrests per 100,000 population

Median age

Proportion males married

Age-specific mortality

Proportion educated

Proportion government expenditures by type

Contextual Data

Additional Wave 3 and 4 tract data

Supplementary tract-level databases (some county/state level variables as well)

Includes transportation and commuting measures including rural-urban commuting area codes and the extent of major roadways, climate descriptors, the presence of particular amenities, and state-level tobacco control influences.

UNC CAROLINA POPULATION CENTER

Wave I and III ONE (Contextual) Data

Obesity & Neighborhood Environment Database.

Linked area-level (5 mile radius) data to the individual data that include community-level measures such as recreation facilities (public, private), transportation options, crime, land use, air pollution, walkability, climate, and cost of living.

UNC CAROLINA POPULATION CENTER

Add Health
The National Longitudinal Study of Adolescent to Adult Health

Contextual Data

Contextual variables may be used like any other variable in models.

As long as you correct for clustering at the school level, you do not need to correct for clustering at any other geographic level.

Contextual variables may be used in a cross-sectional model, a panel model, or a longitudinal model.

In longitudinal models, contextual variables may be time-varying.

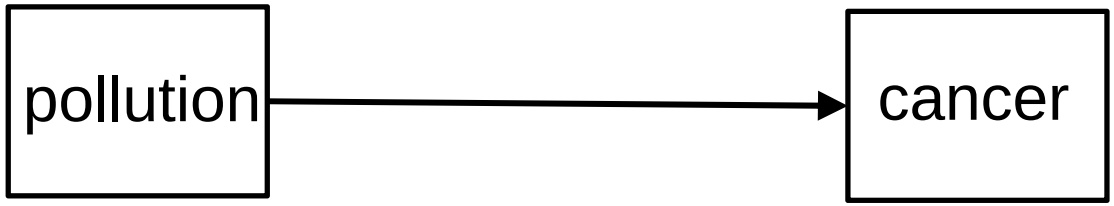
UNC CAROLINA POPULATION CENTER

Add Health
The National Longitudinal Study of Adolescent to Adult Health

Contextual Data

Common Contextual Hypotheses:

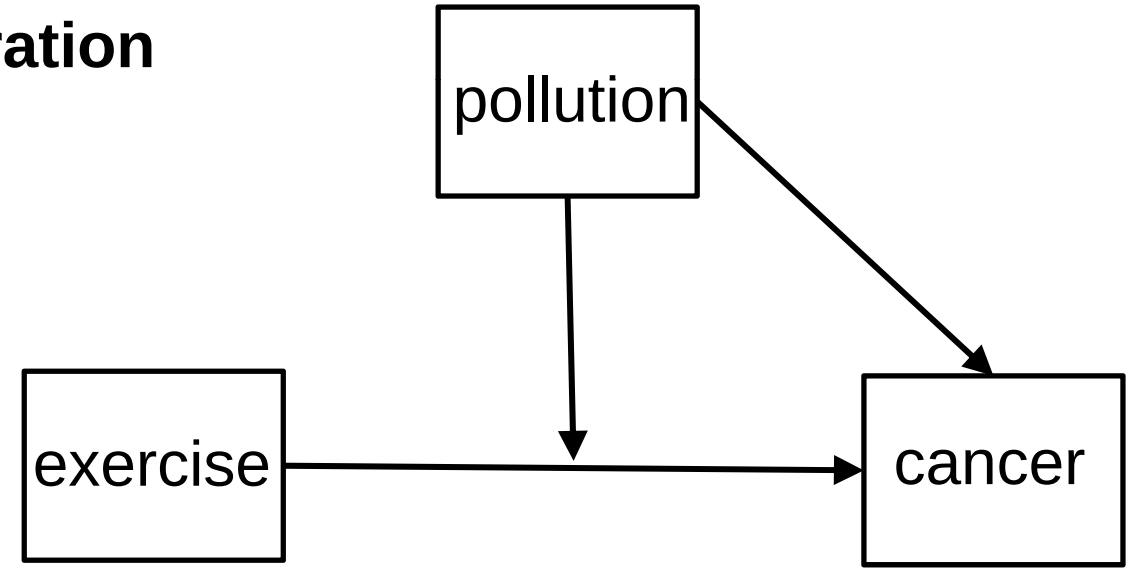
direct influences



Contextual Data

Common Contextual Hypotheses:

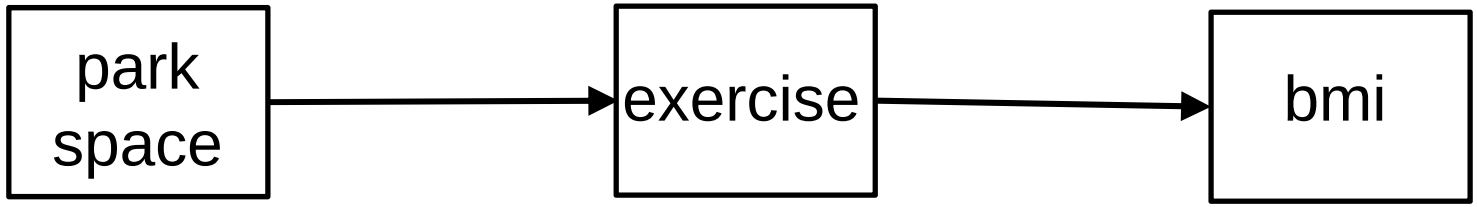
moderation



Contextual Data

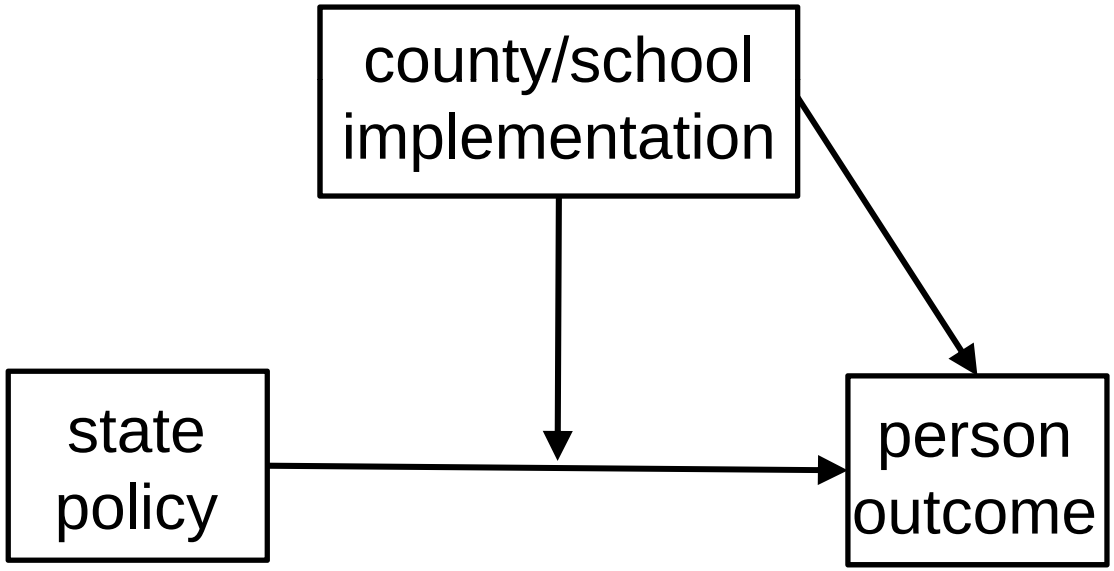
Common Contextual Hypotheses:

Indirect influences



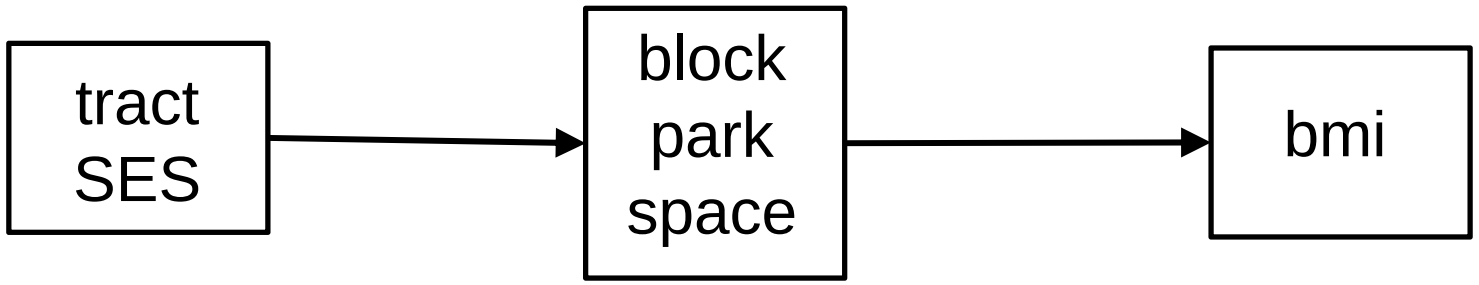
Contextual Data

Multiple Contexts Hypotheses:



Contextual Data

Multiple Contexts Hypotheses:



Contextual Data

Recommendations:

Use the marginal model

- treats nesting at the school level as a nuisance
- uses more efficient single-level weights
- however, no random effects (variance components) estimates

Always use the marginal model for longitudinal (repeated measures, stacked data, three-levels of nesting)

UNC CAROLINA POPULATION CENTER

Contextual Data

Recommendations:

If you use a two-level, multi-level model

- nesting is still at the school level
- contextual variables are fixed effects
- use centering to understand how contextual variables impact variance at the school and person levels