Investigation of Ways of Handling Sampling Weights for Multilevel Model Analyses Using Add Health

Tianji Cai

University of North Carolina

Outlines

- The multilevel model
- Survey data & the role of the sampling weight
- Two approaches for survey data analysis: Design-based vs. Model-based
- Estimating the multilevel model

  Design-based vs. Model-based
- Examples

Multilevel model

- Linear multilevel model

$$\mathbf{Y_i} = \mathbf{X_i}\boldsymbol{\beta} + \mathbf{Z_i}\mathbf{b_i} + \mathbf{e_i}$$

- Joint likelihood function

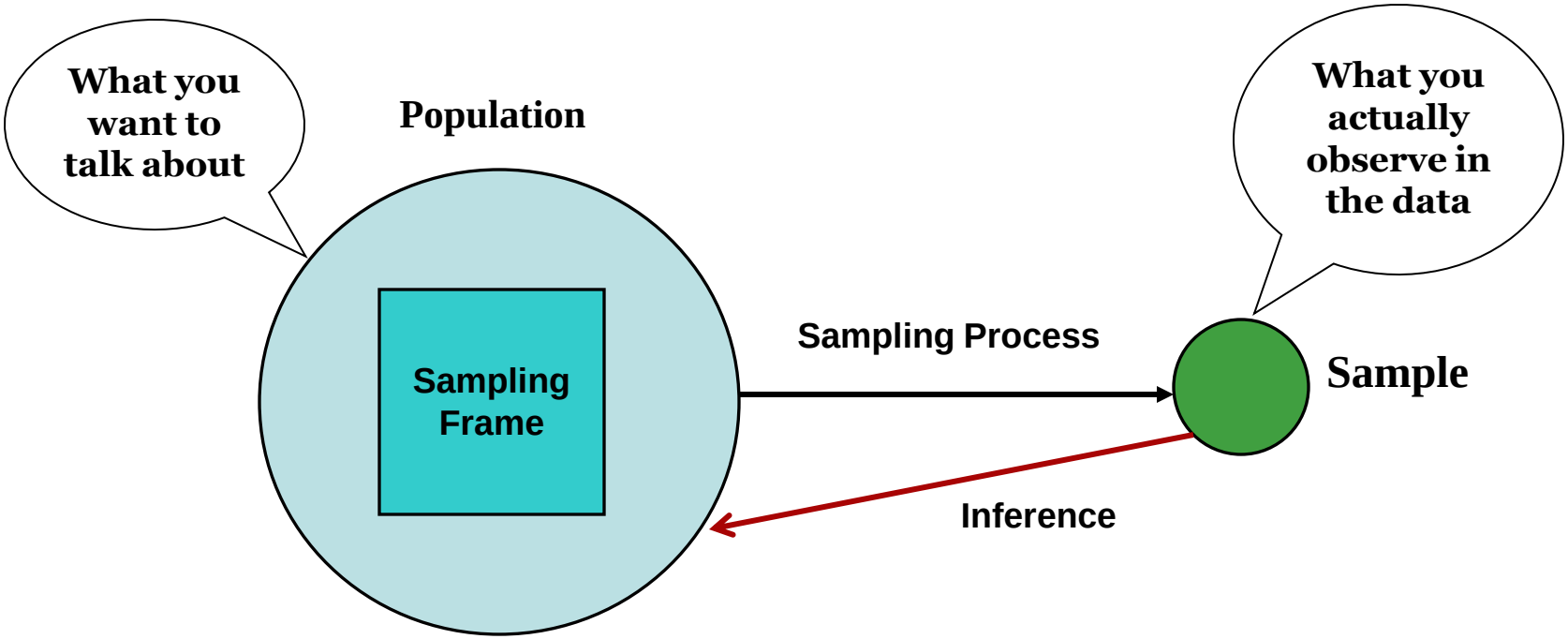
$$L(\mathbf{Y} \mid \mathbf{X}, \mathbf{Z}, \boldsymbol{\beta}, \mathbf{D}, \sigma_e^2) = \frac{\exp\ -\frac{1}{2}\ \mathbf{Y - X\beta}\ '\ \mathbf{V}^{-1}(\mathbf{Y - X\beta})}{2\pi\ ^{\frac{N}{2}}\ |\mathbf{V}|^{\frac{1}{2}}}$$

- Log likelihood function

$$l = \log L(\mathbf{Y} \mid \mathbf{X}, \mathbf{Z}, \boldsymbol{\beta}, \mathbf{D}, \sigma_\varepsilon^2)$$

$$= -\tfrac{1}{2} N \log(2\pi) - \tfrac{1}{2}\log|\mathbf{V}| - \tfrac{1}{2}\ \mathbf{Y - X\beta}\ '\ \mathbf{V}^{-1}(\mathbf{Y - X\beta})$$

What is survey sampling?

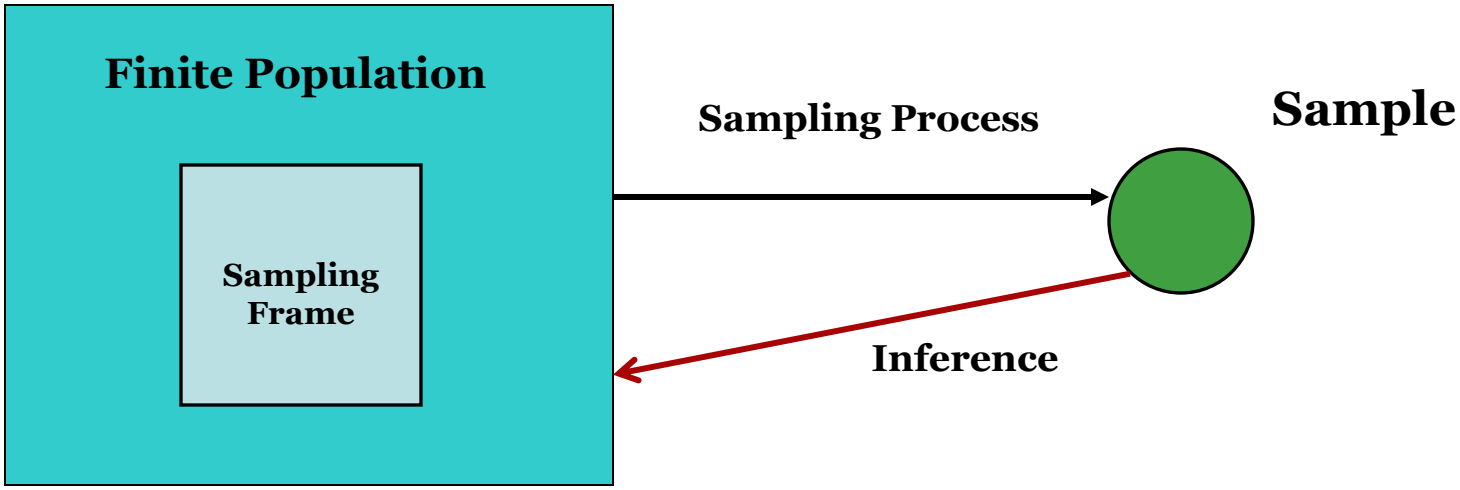


Using data to say something (*make an **inference***) w/ confidence about a whole (population) based on the study of a only a few (sample).

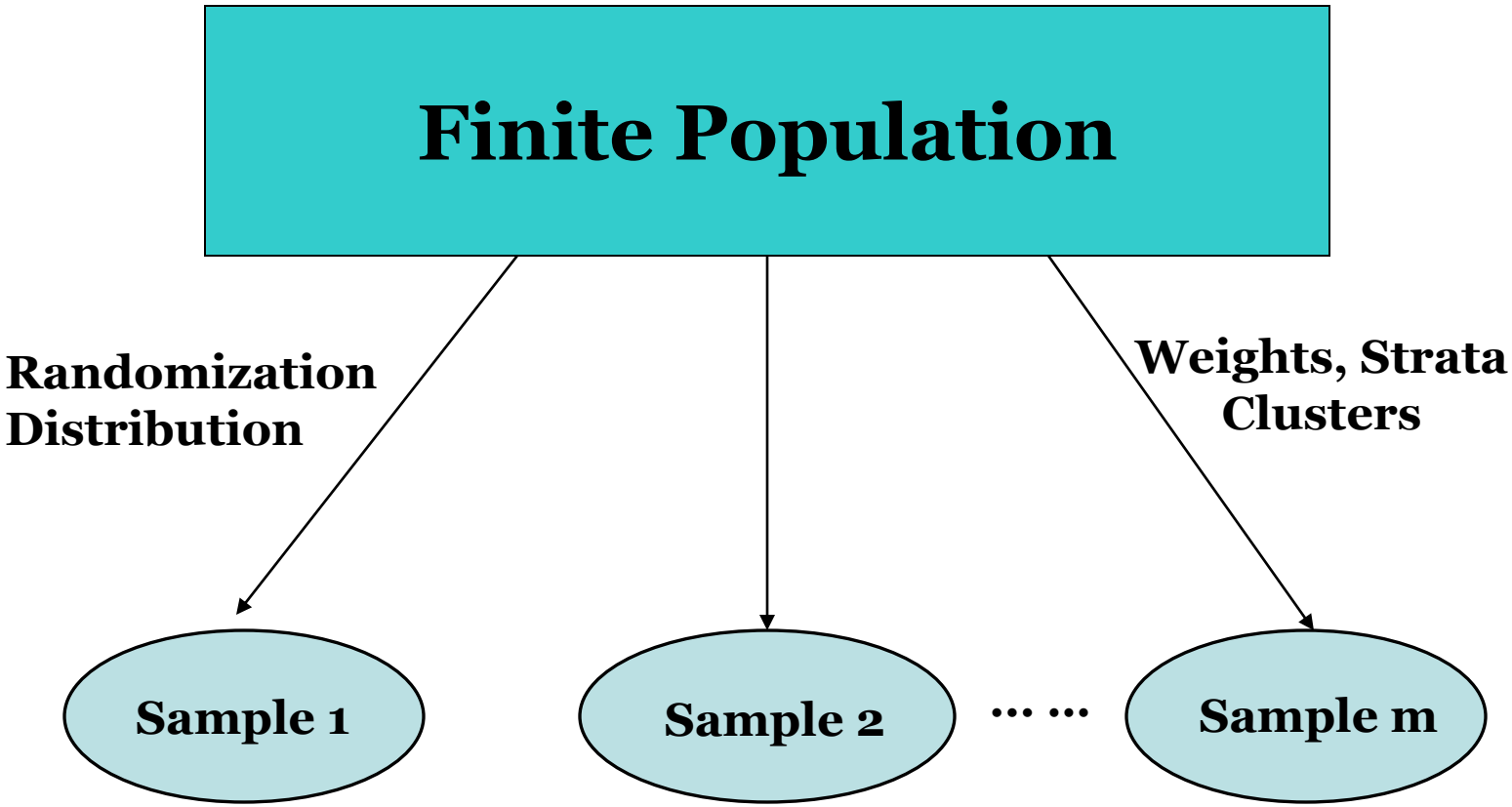
Survey data

- The classical statistical theory assumes simple random sample (SRS) to make inference.
- Most survey data are not SRS.
- The dependence /bias might be introduced by the sampling design, e.g. stratification, clustering, unequal inclusion probability.
- How could we model survey data?
- Two approaches: Design-based vs. Model-based

Design-based approach



Parameters of interest: finite population quantities, eg. mean, total, ratio, and etc.



- Inference: Finite population.
- Variability: Sampling distribution
- Sample design: Strata, Clusters

The role of sampling weight

- To account for the selection probability.
- Sample weights may be adjusted for imbalance due to:
  - *Nonresponse*
  - *Poor frame coverage*
  - *Randomization in sample selection*
- The weight adjustment is usually a multiplicative factor by which the unadjusted weight is multiplied.
- The weight is a random variable!

Design-based approach

- **Probability Weighting**

$$T = \sum_{i=1}^{N} Y_i \rightarrow \hat{T} = \sum_{i=1}^{n} w_i y_i$$

$$\bar{Y} = \sum_{i=1}^{N} Y_i \Big/ N \rightarrow \hat{\bar{Y}} = \sum_{i=1}^{n} w_i y_i \Big/ \sum_{i=1}^{n} w_i$$

- **Pseudo Likelihood Estimation**

$$l(\theta \mid Y) = \sum_{i=1}^{N} \log f(Y_i, \theta) \rightarrow \begin{cases} Y = \beta_0 + \beta_1 X + e \\ \hat{\beta}_{ols} = (X^t X)^{-1} X^t Y \end{cases}$$

$$l(\theta \mid y) = \sum_{i=1}^{n} w_i \log f(y_i, \theta) \rightarrow \hat{\beta}_{ple} = (x^t w x)^{-1} x^t w y$$

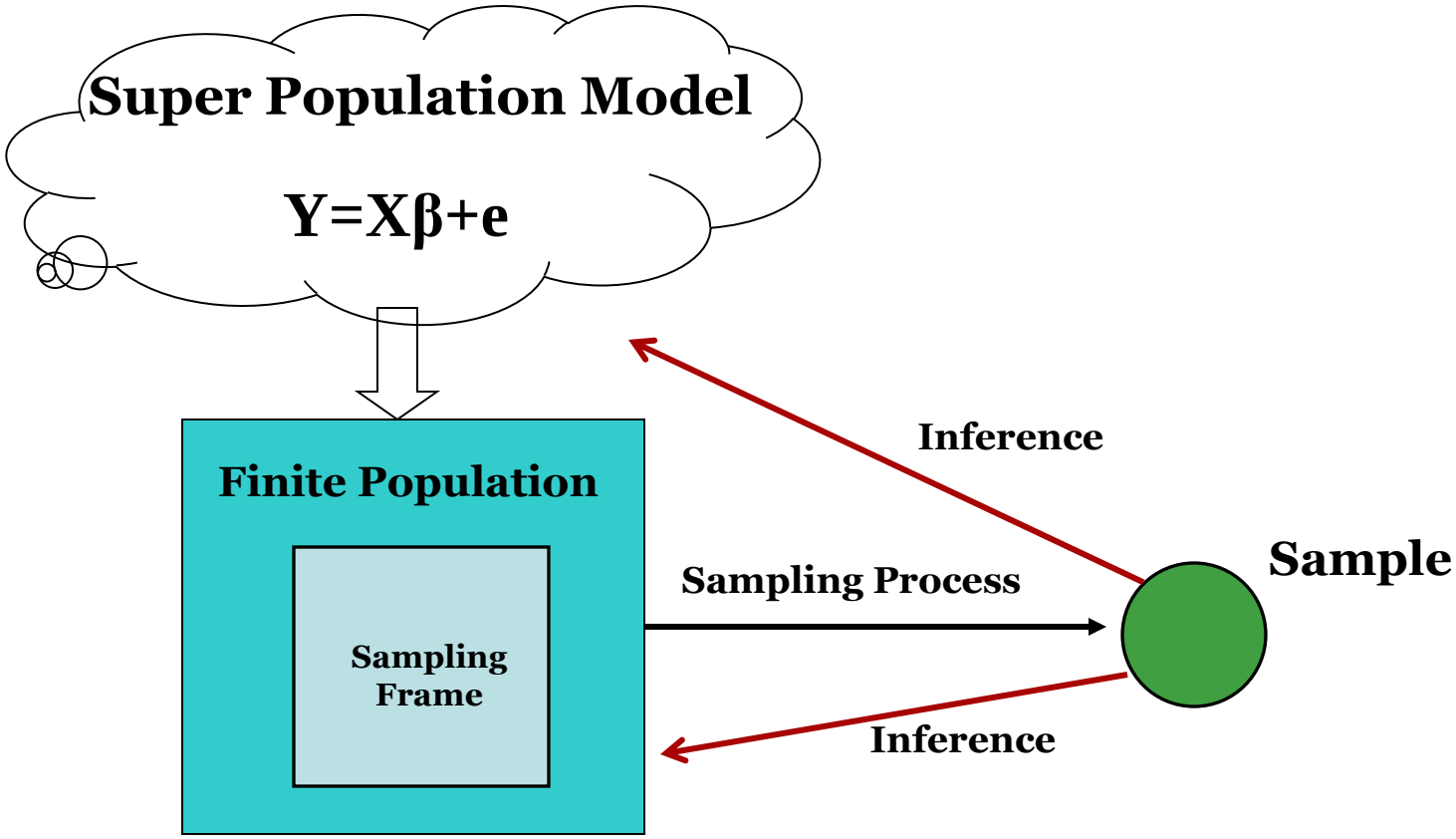
Problems for the design-based approach

- The sampling weight usually is a fixed number for each element in a sample.
- It does not reflect the correlation between the response and the design variables.
- Should we use the same set of weights for all analysis?
- Not available for some models (e.g. Propensity Score Matching), or subsets of the sample (e.g. non-covered).

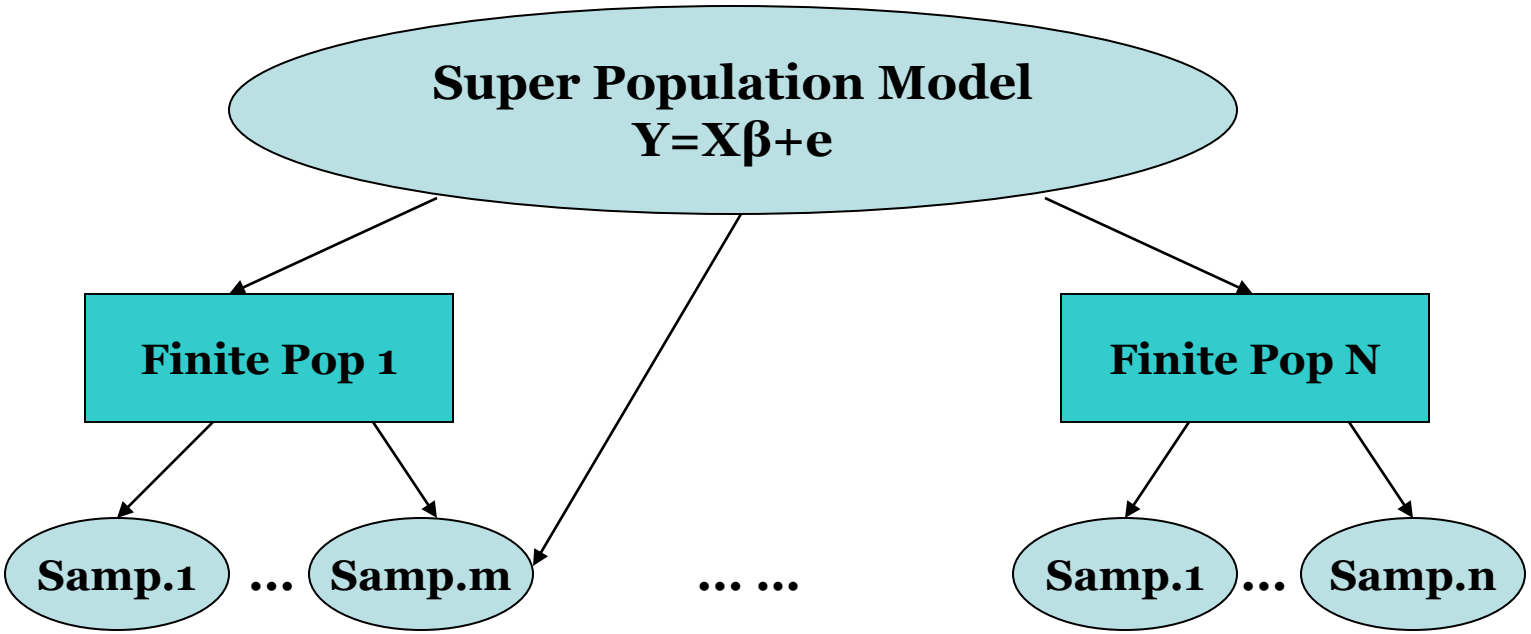
Summary of Design-based approach

- The population elements under study are fixed. Finite population with fixed members.
- The variability is design induced, and from the differences across samples.
- If census were taken, parameters would be known exactly.
- Model is not necessary, e.g. for finite population quantities.
- Weights are needed.

Model-based approach



Target of inference: model parameters β



- Inference: model parameters **β**
- Variability: 1. defined by stochastic model
  2. sampling distribution.
- Sample Design: identifies dependencies

Model-based approach

- The super population elements under study are random.
- The variability is from the model error term and the sampling distribution.
- The design effect can be adjusted.
- If census were taken, parameters might not be known exactly.
- The super population model which describes the underlying stochastic process is necessary.
- The issue of weighting is controversial.

Why the issue of weighting is controversial?

- The sampling design is not correlated to the response.

  e.g. $\pi_i = \Pr(I_i = 1 | x_i, z_i)$

  $\pi_i = \Pr(I_i = 1 | z_i)$

- If the sampling design is correlated to the response.

  e.g. $\pi_i = \Pr(I_i = 1 | y_i, x_i, z_i)$

  $\pi_i = \Pr(I_i = 1 | y_i, x_i)$

  $\pi_i = \Pr(I_i = 1 | y_i)$

Estimating the multilevel model Design-based approaches

$$L(\theta_1,\theta_2)=\prod_i\left(\int\left(\prod_j f\ \ Y_{ij}\left|\mathbf{X}_{ij},b_i,\theta_1\right.\right)\phi\ \ b_i\left|\mathbf{Z}_i,\theta_2\right.\ \ db_i\right)$$

- Multilevel Pseudo Maximum Likelihood (MPML)

$$L_s(\theta_1,\theta_2)=\prod_i\left(\int\left(\prod_j f\ \ y_{ij}\left|\mathbf{x}_{ij},b_i,\theta_1\right.^{w_{j|i}}\right)\phi\ \ b_i\left|\mathbf{z}_i,\theta_2\right.\ \ db_i\right)^{w_i}$$

- Probability Weighted Iterative Generalized Least Square (PWIGLS)

After taking partial derivatives, the population quantities in the score function are replaced by the weighted sample statistics.

Estimating the multilevel model Model-based approach

- **Sample Distribution Method**

  Krieger and Pfeffermann (1992; 1997) proposed a method to extract the model of the sample data as a function of the model holding in the population and the sampling design.

$$f_s(y_i \mid x_i) = f(y_i \mid x_i, I_i = 1)$$

$$= \Pr(I_i = 1, y_i \mid x_i) \times f_p(y_i \mid x_i)$$

$$= \frac{\Pr(I_i = 1 \mid y_i, x_i) \times f_p(y_i \mid x_i)}{\Pr(I_i = 1 \mid x_i)}$$

Sample distribution method

- In general, the probability of inclusion

$$\pi_i = \Pr(I_i = 1 \mid y_i, x_i, z_i) \neq \Pr(I_i = 1 \mid y_i, x_i)$$

Known

Want to know

- However,

$$\Pr(I_i = 1 \mid y_i, x_i)$$

$$= \int \Pr(I_i = 1 \mid y_i, x_i, \pi_i) f_p(\pi_i \mid y_i, x_i) d\pi_i$$

$$= E_p(\pi_i \mid y_i, x_i)$$

Sample distribution method

$$f_s(y_i \mid x_i) = \frac{E_p(\pi_i \mid y_i, x_i) \times f_p(y_i \mid x_i)}{\int E_p(\pi_i \mid y_i, x_i) f_p(y_i \mid x_i) dy_i}$$

- Need to find the form of $E_p(\pi_i \mid y_i, x_i)$
- Two proposals

$$E_p(\pi_i \mid y_i, x_i) \approx \sum_{j=0}^{J} A_j y_i^j + h \ \mathbf{x}_i$$

$$E_p(\pi_i \mid y_i, x_i) \approx \exp\left(\sum_{j=0}^{J} A_j y_i^j + h \ \mathbf{x}_i \right)$$

Sample distribution method : the multilevel model

- Assume exponential approximation for two levels

$$E_p(\pi_i \mid \mu_i, \mathbf{z}_i) \approx g \;\; \mathbf{z}_i \;\; \exp\left(\sum_{r=0}^{R} b_r \mu_i^r \right)$$

$$E_p \;\; \pi_{j|i} \left| y_{ij}, \mathbf{x}_{ij}, \mu_i \;\; \approx k \;\; \mathbf{x}_{ij}, \mu_i \;\; \exp\left(\sum_{h=0}^{H} d_h y_{ij}^h \right) \right.$$

- For simplicity's sake

$$E_p(\pi_i \mid \mu_i, \mathbf{z}_i) \approx \exp \;\; b_1 \mu_i$$

$$E_p \;\; \pi_{j|i} \left| y_{ij}, \mathbf{x}_{ij}, \mu_i \;\; \approx \exp \;\; d_1 y_{ij} \right.$$

Sample distribution method : Multilevel case

- Sample marginal distribution of μ_i

$$f_s\left(\mu_i \middle| \mathbf{z}_i\right) = \frac{1}{\sqrt{2\pi\sigma_\mu^2}} \exp\left(-\frac{\left(\mu_i - \left(b_1\sigma_\mu^2 + \mathbf{z}_i'\gamma\right)\right)^2}{2\sigma_\mu^2} \right)$$

- Sample marginal distribution of y_{ij}

$$f_{s_i}\left(y_{ij} \middle| \mu_i, \mathbf{x}_{ij}\right) = \frac{1}{\sqrt{2\pi\sigma_e^2}} \exp\left(-\frac{\left(y_{ij} - \left(\mu_i + \mathbf{x}_{\mathbf{ij}}'\beta + d_1\sigma_e^2\right)\right)^2}{2\sigma_e^2} \right)$$

Sample distribution method : Multilevel case

- Sample marginal distribution of $\mathbf{y}_i$

$$f_s(\mathbf{y}_i) = 2\pi^{-\frac{m_i}{2}} \sigma_e^{2\,-\frac{m_i-1}{2}} (m_i\sigma_\mu^2 + \sigma_e^2)^{-\frac{1}{2}}$$

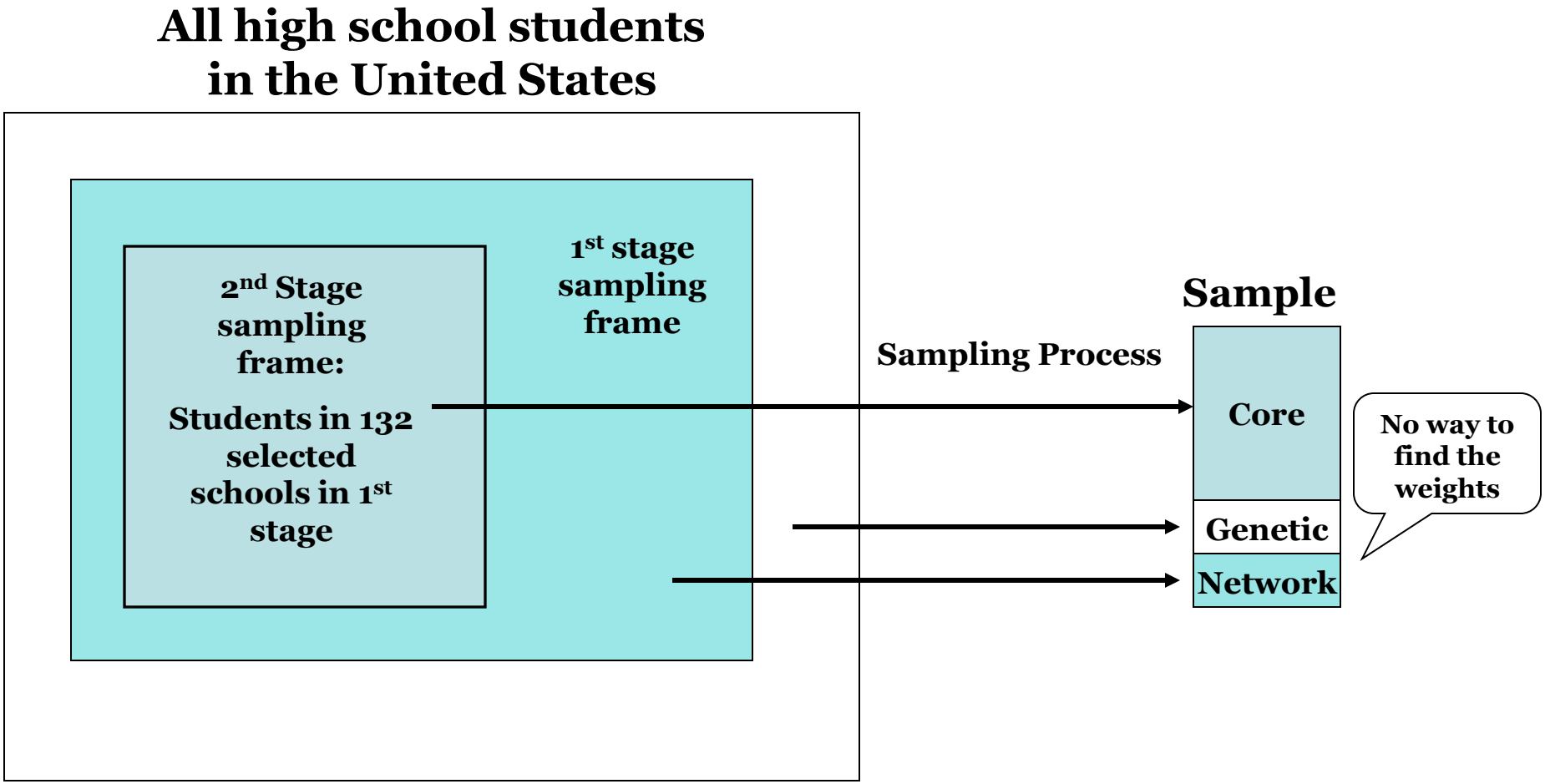
$$\times \exp\left(-\frac{1}{2\sigma_e^2} \sum_{j=1}^{m_i} \left(y_{ij} - (\mathbf{z}_i'\gamma + \mathbf{x}_{ij}'\beta + b_1\sigma_\mu^2 + d_1\sigma_e^2) \right)^2 \right)$$

$$\times \exp\left(\frac{\sigma_\mu^2}{2\sigma_e^2 (m_i\sigma_\mu^2 + \sigma_e^2)} \left(\sum_{j=1}^{m_i} \left(y_{ij} - (\mathbf{z}_i'\gamma + \mathbf{x}_{ij}'\beta + b_1\sigma_\mu^2 + d_1\sigma_e^2) \right) \right)^2 \right)$$

- Joint sample distribution of $\mathbf{y}$

$$f_s = \prod_{i=1}^{m} f_s(\mathbf{y}_i)$$

Add Health example



Cross-sectional multilevel model

- Response variable: a delinquency scale
  Sum of 12 standard items (Guo etc (2008))
- Covariates:
  age at baseline and race
  living with two biological parents (0-1)
  parental unemployment
  education level
  daily family meals
  repeating a grade (0-1)
- Cluster--- school
- 2-level weights: SCHADMWT, & GSWGT1

Estimation

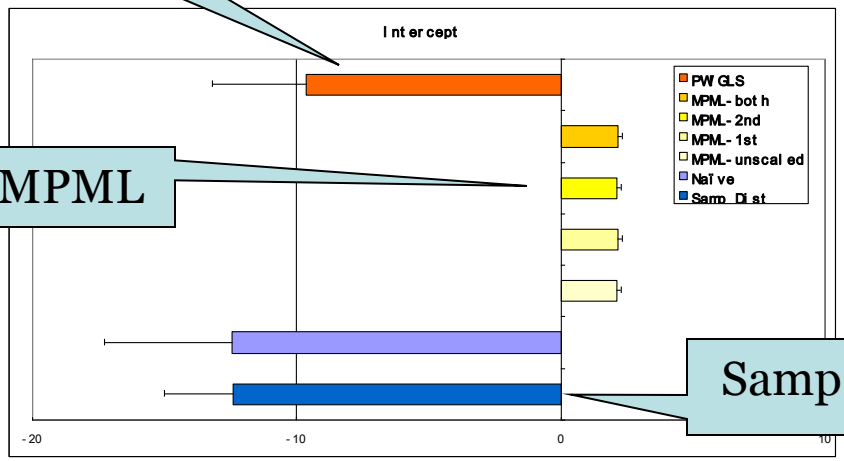
- MPML
  Mplus 5.2 & GLLAMM(STATA)
- PWIGLS
  LISREL 8.7
- The naïve method
  PROC MIXED (SAS 9.2)
- The sample distribution method
  SAS IML

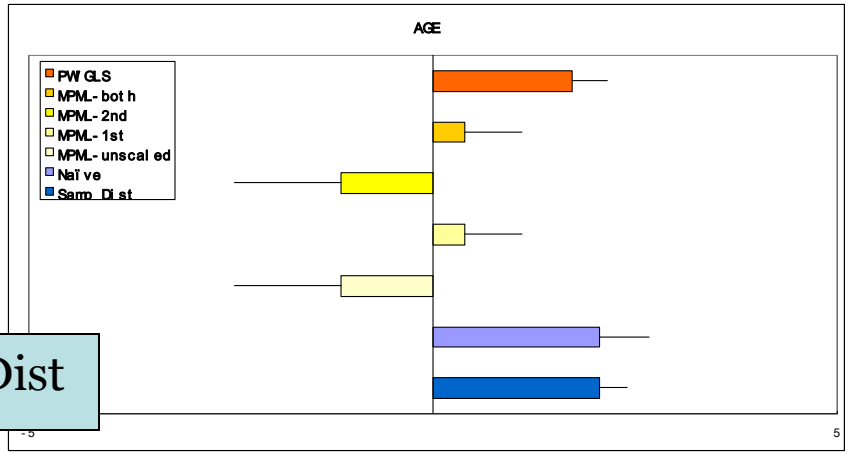
Findings

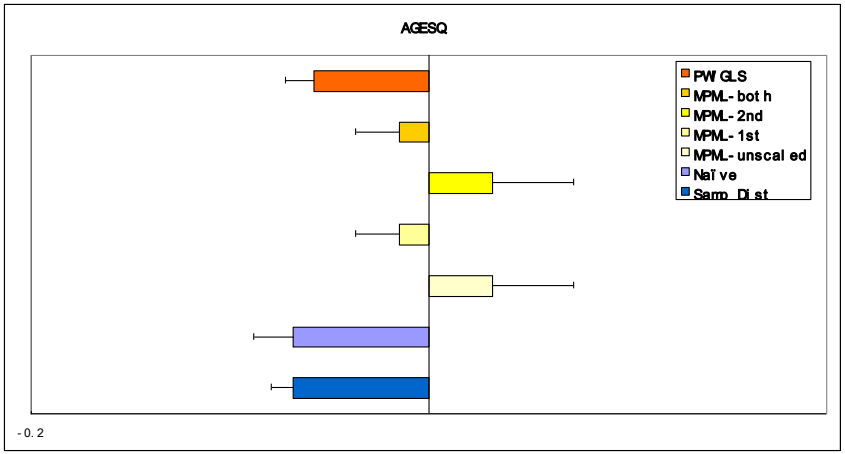
PWIGLS

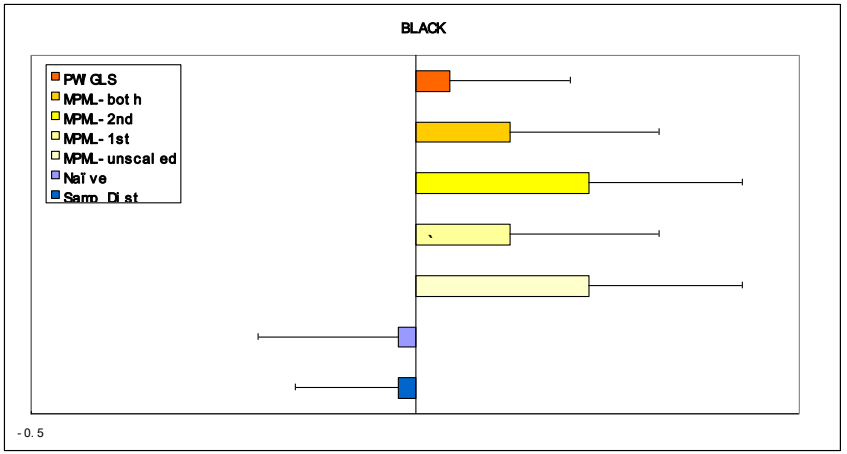
MPML

Samp Dist

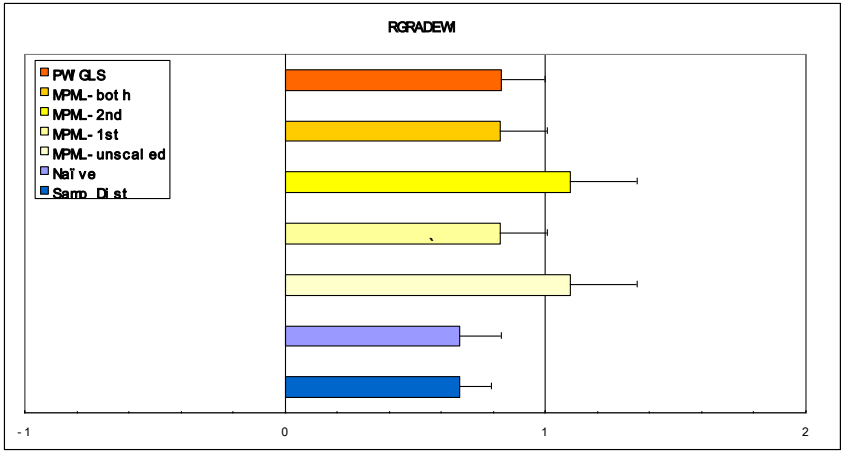


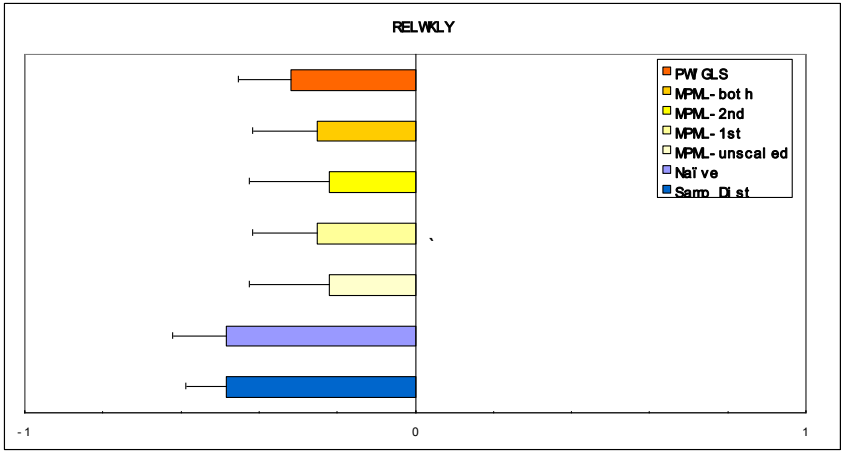


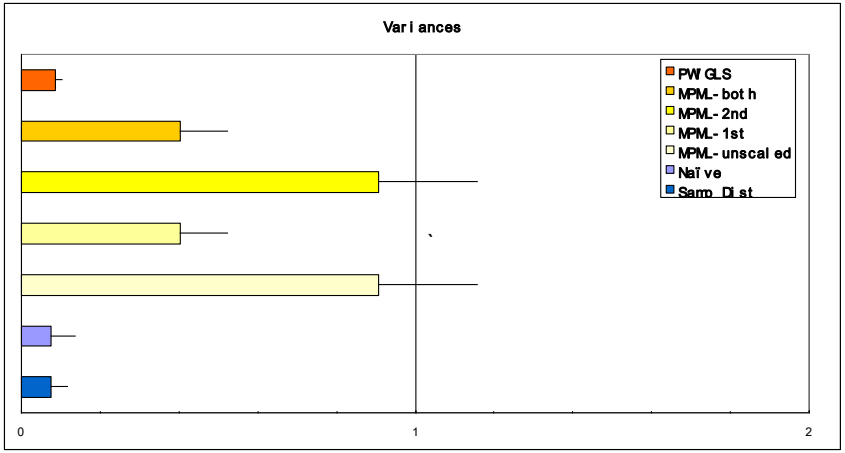


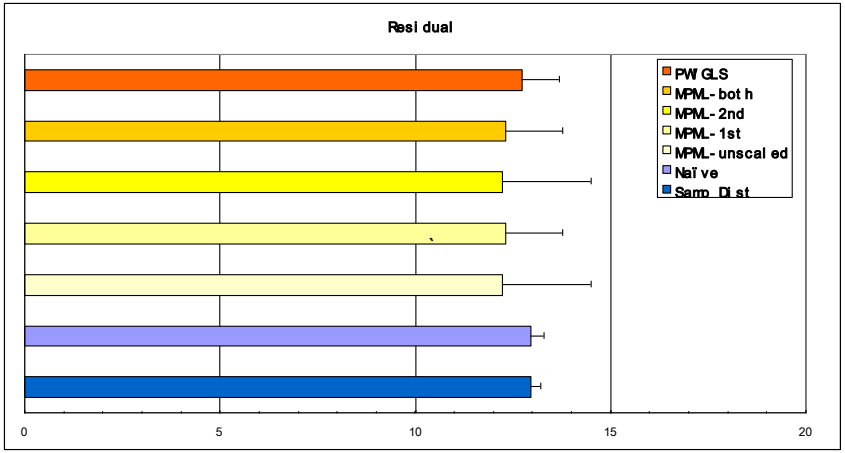


Results


RGRADEWI
PWGLS
MPML- both
MPML- 2nd
MPML- 1st
MPML- unscaled
Naïve
Samp. Dist
-1
0
1
2


RELWKLY
PWGLS
MPML- both
MPML- 2nd
MPML- 1st
MPML- unscaled
Naïve
Samp. Dist
-1
0
1


Variances
PWGLS
MPML- both
MPML- 2nd
MPML- 1st
MPML- unscaled
Naïve
Samp. Dist
0
1
2


Residual
PWGLS
MPML- both
MPML- 2nd
MPML- 1st
MPML- unscaled
Naïve
Samp. Dist
0
5
10
15
20

Summary of findings

- Different weighting methods give very different estimates.
- Different scaling methods also produce different estimates, especially the scaling on the lower level weights.
- The computational algorithms have substantial influences on estimation.

  e.g. GLLAMM did not converge after 16,000 iterations.

Summary of findings

- The naïve method, the sample distribution method, and PWIGLS most of times produce similar estimates, but MPML does not.
- When the sampling weights are not available, the sample distribution method might be a reasonable choice to control the design effects.

Reference

1. Asparouhov, T. (2004). Weighting for Unequal Probability of Selection in Multilevel Modeling. Mplus Web Notes: No.8.

2. Eideh, A. H. and Nathan, G. (2009). Two-Stage Informative Cluster Sampling with application in Small Area Estimation. Journal of Statistical Planning and Inference.139, 3088-3101.

3. Guo, Guang, Michael Roettger, and Tianji Cai. (2008). The Integration of Genetic Propensities into Social Control Models of Delinquency and Violence among Male Youths. American Sociological Review 73:543-568.

4. Pfeffermann, D., Krieger, A.M., and Rinott, Y. (1998). Parametric Distributions of Complex Survey Data Under Informative Probability Sampling. Statistica Sinica, 8, 1087-1114.

5. Rabe-Hesketh, S. and Skrondal, A. (2006), Multilevel Modeling of Complex Survey Data, Journal of the Royal Statistical Society, Series A, 169, 805–827.