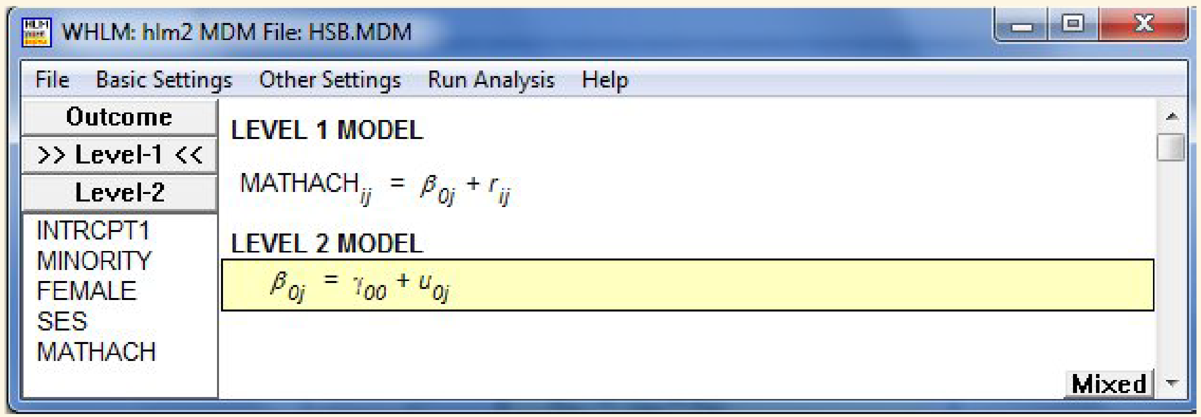
MULTILEVEL MODELING OF COMPLEX SURVEY DATA

Yongchao Ma
ytma@umich.edu

July 22 2024

MULTILEVEL REGRESSION MODEL

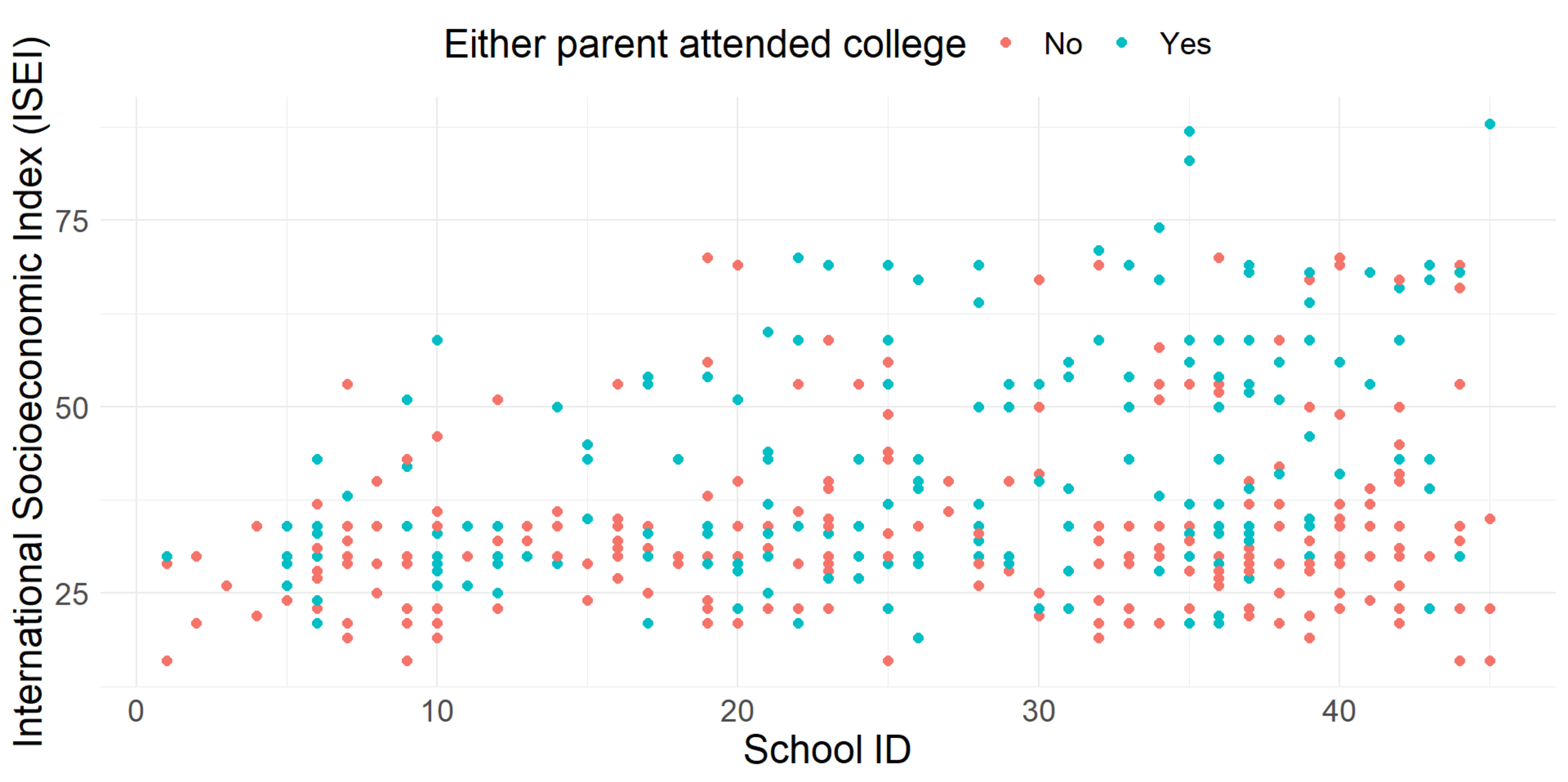
- Known in the literature as
  - Hierarchical linear model (HLM)
  - Random coefficient model
  - Variance components model
  - Mixed model



NESTED DATA STRUCTURE

- Data are often nested in a hierarchical structure
  - Students within schools
  - Patients within hospitals
  - Repeated measures within subjects
  - …

NESTED DATA STRUCTURE



- Are observations independent within schools?

NESTED DATA STRUCTURE

- Observations within the same cluster are often correlated
  - *Intraclass correlation* (ICC) measures the proportion of total variance that is due to between-cluster variance, i.e., the correlation between observations within the same cluster
  - Standard statistical tests are not robust to the violation of independence assumption

- Predictors may exist at different levels
  - Individual-level predictors, e.g., either parent attended college
  - Cluster-level predictors, e.g., average teacher's experience

- Relationship between the outcome and predictors may vary across clusters
  - E.g., the effect of either parent attended college on ISEI may vary across schools

ORDINARY REGRESSION

$$y_i = \beta_0 + \beta_1 x_i + e_i$$

- y_i is the outcome for the i-th observation
- x_i is the predictor for the i-th observation
- β_0 is the intercept
- β_1 is the slope
- ϵ_i is the error term

MULTILEVEL REGRESSION: RANDOM INTERCEPTS

$$y_{ij} = \beta_{0j} + \beta_1 x_{ij} + e_{ij}$$

- y_{ij} is the outcome for the i-th observation in the j-th cluster
- x_{ij} is the predictor for the i-th observation in the j-th cluster
- β_{0j} is the **intercept for the j-th cluster**
- β_1 is the slope
- ϵ_{ij} is the error term

MULTILEVEL REGRESSION: RANDOM SLOPES

$$y_{ij} = \beta_{0j} + \beta_{1j}x_{ij} + e_{ij}$$

- y_{ij} is the outcome for the i-th observation in the j-th cluster
- x_{ij} is the predictor for the i-th observation in the j-th cluster
- β_{0j} is the intercept for the j-th cluster
- β_{ij} is the **slope for the** j-th cluster
- ϵ_{ij} is the error term

MULTILEVEL REGRESSION: RANDOM EFFECTS

- Level 1 (individual level)

$$y_{ij} = \beta_{0j} + \beta_{1j} x_{ij} + e_{ij}$$

- Level 2 (cluster level)

$$\begin{aligned} \beta_{0j} &= \gamma_{00} + u_{0j} \quad \text{where} \quad u_{0j} \sim N(0, \sigma^2_{u0}) \\ \beta_{1j} &= \gamma_{10} + u_{1j} \quad \text{where} \quad u_{1j} \sim N(0, \sigma^2_{u1}) \end{aligned}$$

  - σ^2_{u0} is the variance of random intercepts
  - σ^2_{u1} is the variance of random slopes
- The intercept and slope coefficients are allowed to vary across clusters—*random coefficient* model

MULTILEVEL REGRESSION: PREDICTION AT LEVEL 2

$$y_{ij} = \beta_{0j} + \beta_{1j}x_{ij} + e_{ij}$$
$$\beta_{0j} = \gamma_{00} + \gamma_{01}z_j + u_{0j}$$
$$\beta_{1j} = \gamma_{10} + \gamma_{11}z_j + u_{1j}$$

- z_j is the predictor at the cluster level
- γ_{00} and γ_{01} are the intercept and slope to predict β_{0j}
- γ_{10} and γ_{11} are the intercept and slope to predict β_{1j}
- γ_{00}, γ_{01}, γ_{10}, and γ_{11} are **fixed effects** across clusters
- Between-cluster variation left unexplained by the fixed effects is captured by **random effects** u_{0j} and u_{1j}

MULTILEVEL REGRESSION: MIXED MODEL

- Level 1 (individual level)

$$y_{ij} = \beta_{0j} + \beta_{1j}x_{ij} + e_{ij}$$

- Level 2 (cluster level)

$$\beta_{0j} = \gamma_{00} + \gamma_{01}z_j + u_{0j}$$
$$\beta_{1j} = \gamma_{10} + \gamma_{11}z_j + u_{1j}$$

- Mixed model: a combination of fixed and random effects

$$y_{ij} = \underbrace{\gamma_{00} + \gamma_{01}z_j + \gamma_{10}x_{ij} + \gamma_{11}x_{ij}z_j}_{\text{fixed}} + \underbrace{u_{0j} + u_{1j}x_{ij} + e_{ij}}_{\text{random}}$$

MULTILEVEL REGRESSION: MIXED MODEL

- Fixed part is an ordinary regression model

$$y_{ij} = \underbrace{\gamma_{00} + \gamma_{01} z_j + \gamma_{10} x_{ij} + \gamma_{11} x_{ij} z_j}_{\text{fixed}} + \underbrace{u_{0j} + u_{1j} x_{ij} + e_{ij}}_{\text{random}}$$

- Random part consists three error terms

$$\begin{pmatrix} u_{0j} \\ u_{1j} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_{u0}^2 & \sigma_{u01} \\ \sigma_{u01} & \sigma_{u1}^2 \end{pmatrix} \right)$$

$$e_{ij} \sim N(0, \sigma_e^2)$$

MULTILEVEL REGRESSION: COMPLEX SURVEY DATA

- Multilevel modeling is appropriate for analyzing complex sample survey data
  - In the model-based spirit, effects of *randomly sampled* clusters (possibly within strata) are generally treated as **random effects**
  - Effects of strata (*fixed by design*, and not randomly sampled) are generally treated as **fixed effects**
- How to handle *informative* survey weights?

MULTILEVEL REGRESSION: SURVEY WEIGHTS

MODEL-BASED APPROACH

- Include the variables used to build the weights or appropriate functional forms of the weight values themselves as **fixed effects** in the model (Dumouchel and Duncan 1983; Little 1991; Korn and Graubard 1999; Fuller 2009)
  - Only requires standard multilevel modeling software
  - Good model specification is very important

MULTILEVEL REGRESSION: SURVEY WEIGHTS

MODEL-BASED APPROACH

- Example: Estimating a linear model when the underlying data has a quadratic trend (misspecified model)
  - Truth: $y = 2x - x^2 + e \quad \text{where} \quad e \sim N(0,1)$
  - Model: $y = \beta_0 + \beta_1 x + e \quad \text{where} \quad e \sim N(0,\sigma^2)$
  - Selection probability: $\Pr(I=1) \propto x^{0.75}$
  - Target value: linear slope *in population*
    - Population MLE is best linear approximation to underlying quadratic relationship
    - True population intercept: 0.168
    - True population slope: 0.975

	Bias (Unweighted)	**Bias (Weighted)**	**Bias (Weight as covariate)**
Intercept	0.125	0.003	0.337
Slope	-0.089	-0.003	-0.381

MULTILEVEL REGRESSION: SURVEY WEIGHTS

HYBRID APPROACH

- Pseudo maximum likelihood estimation (PMLE): using the weights at *all* levels to compute unbiased estimates of multilevel model parameters (Pfeffermann et al. 1998; Rabe-Hesketh and Skrondal 2006; Carle 2009)
  - The computation of unbiased population estimates provides *some* protection against model misspecification
    - Even if the model is poorly specified, the estimates are still unbiased
- Variance estimation with respect to both sample design AND model (Pfeffermann et al. 1998)
  - Follows Binder's linearization method for implicit estimators that maximize pseudo-likelihood functions
  - Also accounts for stratification and cluster sampling
  - Referred to as "robust" or "sandwich-type" standard errors in software implementing these methods

MULTILEVEL REGRESSION: SURVEY WEIGHTS

HYBRID APPROACH: DATA REQUIREMENTS

- Implementing the hybrid approach requires **weights at each level of the nested data structure**
  - Level-1 weights: inverses of *conditional* probabilities of selection (responding), given that a Level-2 cluster was sampled
  - Level-2 weights: inverses of probabilities of selection for sampling clusters
    - Should be conditional weights if considering a three-level model
- Also need cluster codes and stratum codes

MULTILEVEL REGRESSION: SURVEY WEIGHTS

HYBRID APPROACH: DATA REQUIREMENTS

> ⚠ DO NOT use the overall (adjusted) sampling weights that are typically provided in public-use data files

- The overall inclusion probabilities do not carry forward sufficient information for bias correction when estimating multilevel models
  - The simple design-based idea of pseudo-likelihood (weighting likelihood contributions by overall sampling weight, and *assuming independent observations* in finite population) is not sufficient, generally due to **random effects** in the model-based approach

- Separate *conditional* weights are needed at lower levels to specify the appropriate likelihood function, and compute unbiased estimates of both fixed-effect and covariance parameters
  - We need to strip out the probability that a higher-level cluster was sampled from the overall weights that are usually provided for respondents

MULTILEVEL REGRESSION: SURVEY WEIGHTS

HYBRID APPROACH: WEIGHT SCALING

- Sums of *conditional* weights at Level 1 over-state the actual sample sizes within clusters
  - Consider scaling the conditional weights at Level 1 (and other lower levels) of the data hierarchy, especially when the sampling is non-informative
  - A failure to do so can lead to bias in parameter estimates, especially for small samples (Pfeffermann et al. 1998)
  - Weight scaling is particularly important for multilevel logistic regression models(Rabe-Hesketh and Skrondal 2006)
  - **Consistent in the literature**: it is better to scale the weights than do nothing when estimating multilevel models
- Many methods for scaling the weights have been proposed in the literature, with no consistent "winners" in terms of bias reduction (Carle 2009)

MULTILEVEL REGRESSION: SURVEY WEIGHTS

HYBRID APPROACH: WEIGHT SCALING

- Two methods used most often, and resulting in the least bias in estimates based on simulation studies
  - Method 1 scales weights by "design effects" to yield effective sample sizes within clusters, rather than "naïve" (or nominal) sample sizes
  - Method 2 scales weights so that they sum to actual sample sizes rather than weighted sample sizes
    - Good for informative weights

MULTILEVEL REGRESSION: SURVEY WEIGHTS

HYBRID APPROACH: WEIGHT SCALING

> Recommendation: consider the sensitivity of inferences to all available methods of weight scaling

- If no notable differences, use Method 2 (size)
  - Good for point estimates and estimates of variance components
  - Simulations reported by Rabe-Hesketh and Skrondal (2006) also support the use of Method 2 (size) for multilevel logistic regression models
- If differences are observed (rare), Carle (2009) suggests the use of Method 1 (effective), where there is more of a focus on the variance component estimates

MULTILEVEL REGRESSION: SURVEY WEIGHTS

HYBRID APPROACH: SOFTWARE IMPLEMENTATION

- Stata (three scaling options)
  - `gllamm` in Stata (manual scaling needed)
- R `svylme` package (for continuous DVs only) (Lumley and Huang 2024)
  - Instead of PML, uses a pairwise composite likelihood approach
  - No large-cluster assumption needed and no weight scaling
  - Inefficient for estimating variance components, especially for random-intercept variance
- SAS (`PROC GLIMMIX`)
- Mplus (several alternative scaling methods)
- HLM (automatic weight scaling using the “size” method)
- MLwiN (automatic weight scaling)

EXAMPLE: PISA DATA (2000)

- The Programme for International Student Assessment (PISA) is a worldwide study by the Organisation for Economic Co-operation and Development (OECD) in member and non-member nations
- Dependent variable:
  - ISEI (International Socio-Economic Index) of the student
- Predictors:
  - COLLEGE: Indicator of whether the highest level of education for either parent is college
- Design variables:
  - ID_SCHOOL: Code for randomly sampled school (cluster)
  - W_FSTUWT: Overall final student weight (NOT conditional)
  - WNRSCHBW: Final school weight (this is the weight that is rarely provided)

EXAMPLE: PISA DATA (2000)

MODEL SPECIFICATION IN R: UNWEIGHTED

```r
# Uncorrelated random effects are specified with the double-bar notation
lme4::lmer(isei ~ college + (college || id_school),
           REML = TRUE,
           data = pisa)
```

```
Linear mixed model fit by REML ['lmerMod']
Formula: isei ~ college + ((1 | id_school) + (0 + college | id_school))
   Data: pisa
REML criterion at convergence: 17220.36
Random effects:
 Groups      Name        Std.Dev. Corr
 id_school   (Intercept)  1.110
 id_school.1 collegeNo    3.354
             collegeYes   7.814   0.72
 Residual                14.851
Number of obs: 2069, groups:  id_school, 148
Fixed Effects:
(Intercept)   collegeYes
      38.77        12.60
optimizer (nloptwrap) convergence code: 0 (OK) ; 0 optimizer warnings; 2 lme4
warnings
```

EXAMPLE: PISA DATA (2000)

MODEL SPECIFICATION IN STATA: UNWEIGHTED

```
mixed isei college || id_school: college, ///
covariance(independent) variance

                                                    Wald chi2(1)        =     195.93
Log likelihood = -8611.8768                         Prob > chi2         =     0.0000

------------------------------------------------------------------------------
        isei |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     college |   12.64623   .9034645    14.00   0.000     10.87548    14.41699
       _cons |   38.78531    .619744    62.58   0.000     37.57064    39.99999
------------------------------------------------------------------------------

------------------------------------------------------------------------------
  Random-effects Parameters  |   Estimate   Std. Err.     [95% Conf. Interval]
```

EXAMPLE: PISA DATA (2000)

MODEL SPECIFICATION IN R: FINAL OVERALL WEIGHTS AS COVARIATES

```
# Uncorrelated random effects are specified with the double-bar notation
lme4::lmer(isei ~ college + w_fstuwt + (college || id_school),
           REML = TRUE,
           data = pisa)
```

```
Linear mixed model fit by REML ['lmerMod']
Formula: isei ~ college + w_fstuwt + ((1 | id_school) + (0 + college |
    id_school))
   Data: pisa
REML criterion at convergence: 17228.44
Random effects:
 Groups      Name        Std.Dev. Corr
 id_school   (Intercept)  0.606
 id_school.1 collegeNo    3.275
             collegeYes   7.656   0.69
 Residual                14.867
Number of obs: 2069, groups:  id_school, 148
Fixed Effects:
(Intercept)   collegeYes     w_fstuwt
  36.897853    12.554921     0.002294
```

EXAMPLE: PISA DATA (2000)

MODEL SPECIFICATION IN STATA: FINAL OVERALL WEIGHTS AS COVARIATES

```
mixed isei college w_fstuwt || id_school: college, ///
covariance(independent) variance

                                                   Wald chi2(2)      =    203.31
Log likelihood = -8609.9137                        Prob > chi2       =    0.0000

------------------------------------------------------------------------------
        isei |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     college |   12.59723   .9003722    13.99   0.000     10.83254    14.36193
    w_fstuwt |   .0024003   .0011771     2.04   0.041     .0000933    .0047073
       _cons |   36.82151   1.135727    32.42   0.000     34.59553    39.04749
------------------------------------------------------------------------------

------------------------------------------------------------------------------
  Random-effects Parameters  |   Estimate   Std. Err.     [95% Conf. Interval]
```

EXAMPLE: PISA DATA (2000)

MODEL SPECIFICATION IN R: WEIGHTED PAIRWISE COMPOSITE LIKELIHOOD APPROACH

```r
# Compute the conditional student weights
pisa$w_condstuwt <- with(pisa, w_fstuwt / wnrschbw)

# Assign a unique ID to each student
pisa$id_student <- 1:nrow(pisa)

# Specify the survey design
dpisa <- survey::svydesign(
  id = ~ id_school + id_student,
  weight = ~ wnrschbw + w_condstuwt,
  data = pisa
)
```

EXAMPLE: PISA DATA (2000)

MODEL SPECIFICATION IN R: WEIGHTED PAIRWISE COMPOSITE LIKELIHOOD APPROACH

```r
# Uncorrelated random effects are specified with the double-bar notation
svylme::svy2lme(isei ~ college + (college || id_school), design = dpisa)
```

```
Linear mixed model fitted by pairwise pseudolikelihood
Formula: isei ~ college + (college || id_school)
Random effects:
                       Std.Dev.
id_school:(Intercept)    0.0000
id_school1:collegeNo     0.5617
id_school2:collegeYes    7.4402
Residual:     14.7629
 Fixed effects:
               beta      SE      t       p
(Intercept) 40.8120  0.8005 50.98 <2e-16
collegeYes  15.9528  1.4417 11.07 <2e-16
```

EXAMPLE: PISA DATA (2000)

MODEL SPECIFICATION IN STATA: WEIGHTED, SCALING METHOD 1 (EFFECTIVE)

```
gen conwt = w_fstuwt / wnrschbw

mixed isei college [pw = conwt] || id_school: college, ///
covariance(independent) variance pweight(wnrschbw) pwscale(effective)

                                                   Wald chi2(1)       =    100.84
Log pseudolikelihood = -1439307.8                  Prob > chi2        =    0.0000

                             (Std. Err. adjusted for 148 clusters in id_school)
------------------------------------------------------------------------------
             |               Robust
        isei |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     college |   14.28032   1.422044    10.04   0.000     11.49316    17.06747
       _cons |   35.88949   .9100379    39.44   0.000     34.10585    37.67313
```

EXAMPLE: PISA DATA (2000)

MODEL SPECIFICATION IN STATA: WEIGHTED, SCALING METHOD 2 (SIZE)

```
mixed isei college [pw = conwt] || id_school: college, ///
covariance(independent) variance pweight(wnrschbw) pwscale(size)

                                               Wald chi2(1)       =     100.87
Log pseudolikelihood =   -1443258              Prob > chi2        =     0.0000

                            (Std. Err. adjusted for 148 clusters in id_school)
------------------------------------------------------------------------------
             |               Robust
        isei |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     college |   14.27681   1.421497    10.04   0.000     11.49073     17.0629
       _cons |   35.89078   .9099169    39.44   0.000     34.10738    37.67419
------------------------------------------------------------------------------
```

EXAMPLE: PISA DATA (2000)

SUMMARY OF RESULTS IN STATA

Parameter	No Weights	Weights as a Covariate	Scaling Method 1: Effective	Scaling Method 2: Size
Fixed Effects				
Intercept	38.79 (0.62)	36.82 (1.14)	35.89 (0.91)	35.89 (0.91)
COLLEGE	12.65 (0.90)	12.60 (0.90)	14.28 (1.42)	14.28 (1.42)
WEIGHT		<0.01 (<0.01)		
Variance Components				
Var(Intercepts)	16.14 (5.45)	13.94 (5.22)	17.74 (6.43)	17.79 (6.43)
Var(College)	43.12 (10.85)	42.26 (10.58)	41.03 (13.74)	41.06 (13.73)
Var(Residuals)	219.33 (7.20)	219.92 (7.22)	214.96 (12.82)	214.92 (12.84)
Pseudo Log(L)	–8,611.88	–8,609.91	–1,439,307.8	–1,443,258.0

EXAMPLE: PISA DATA (2000)

SUMMARY OF RESULTS IN STATA

- Parental college education has a strong effect on SES, regardless of the method used
- Weighted estimates of fixed effects are different from unweighted estimates, especially the intercept (i.e., mean for students with non-college educated parents)
- Including the student-level weight as a covariate changes interpretations of parameters

EXAMPLE: PISA DATA (2000)

SUMMARY OF RESULTS IN STATA

- Weighted estimates of variance components differ
  - More evidence of variability across schools in means for students with non-college educated parents (i.e., the random intercepts) when computing weighted estimates
  - Less variability in college vs. non-college gaps (i.e., the random coefficients) across the sampled schools
- Weight scaling methods do not result in different estimates or conclusions
  - use Method 2 (size)
- Robust standard errors for weighted estimates are generally larger, but do not change inferences

FINAL POINTS

- Survey agencies generally do not release the weight information necessary to implement the “hybrid” multilevel modeling approaches (mainly the weights associated with sampling clusters)
- Analysts thus need to resort to the model-based approach, which can be problematic if weights are informative and not accounted for
- Software is not widely available for the “hybrid” approach, but this approach is best at reducing bias
- Make sure that a multilevel model is what you need for your research objectives (e.g., interest in variance components, interest in cross-level interactions, etc.)

REFERENCES

Carle, Adam C. 2009. "Fitting Multilevel Models in Complex Survey Data with Design Weights: Recommendations." *BMC Medical Research Methodology* 9 (1): 49. https://doi.org/10.1186/1471-2288-9-49.

Dumouchel, William H., and Greg J. Duncan. 1983. "Using Sample Survey Weights in Multiple Regression Analyses of Stratified Samples." *Journal of the American Statistical Association* 78 (383): 535–43. https://doi.org/10.1080/01621459.1983.10478006.

Fuller, Wayne A. 2009. *Sampling Statistics*. 1st ed. Wiley. https://doi.org/10.1002/9780470523551.

Korn, Edward L., and Barry I. Graubard. 1999. *Analysis of Health Surveys*. 1st ed. Wiley. https://doi.org/10.1002/9781118032619.

Little, Roderick J. 1991. "Inference with Survey Weights." *Journal of Official Statistics* 7: 405–24.

Lumley, Thomas, and Xudong Huang. 2024. "Linear Mixed Models for Complex Survey Data: Implementing and Evaluating Pairwise Likelihood." *Stat* 13 (1): e657. https://doi.org/10.1002/sta4.657.

Pfeffermann, D., C. J. Skinner, D. J. Holmes, H. Goldstein, and J. Rasbash. 1998. "Weighting for Unequal Selection Probabilities in Multilevel Models." *Journal of the Royal Statistical Society Series B: Statistical Methodology* 60 (1): 23–40. https://doi.org/10.1111/1467-9868.00106.

Rabe-Hesketh, Sophia, and Anders Skrondal. 2006. "Multilevel Modelling of Complex Survey Data." *Journal of the Royal Statistical Society Series A: Statistics in Society* 169 (4): 805–27. https://doi.org/10.1111/j.1467-985X.2006.00426.x.