Multilevel Modeling of Complex Survey Data

Sophia Rabe-Hesketh, University of California, Berkeley
and Institute of Education, University of London

Joint work with Anders Skrondal, London School of Economics

2007 West Coast Stata Users Group Meeting
Marina del Rey, October 2007

Outline

- Model-based and design based inference
- Multilevel models and pseudolikelihood
- Pseudo maximum likelihood estimation for U.S. PISA 2000 data
- Scaling of level-1 weights
- Simulation study
- Conclusions

Multistage sampling: U.S. PISA 2000 data

- Program for International Student Assessment (PISA):
  Assess and compare 15 year old students' reading, math, etc.
- Three-stage survey with different probabilities of selection
  - Stage 1: Geographic areas k sampled
  - Stage 2: Schools $j=1,\ldots,n^{(2)}$ sampled with different probabilities π_j (taking into account school non-response)
  - Stage 3: Students $i=1,\ldots,n_j^{(1)}$ sampled from school j, with conditional probabilities $\pi_{i|j}$
- Probability that student i from school j is sampled:

$$\pi_{ij} = \pi_{i|j}\pi_j$$

Model-based and design-based inference

- **Model-based inference**: Target of inference is parameter β in infinite population (parameter of data generating mechanism or statistical model) called **superpopulation** parameter
  - Consistent estimator (assuming simple random sampling) such as maximum likelihood estimator (MLE) yields estimate $\widehat{\beta}$
- **Design-based inference**: Target of inference is statistic in **finite population** (FP), e.g., mean score $\overline{y}^{\mathrm{FP}}$ of all 15-year olds in LA
  - Student who had a $\pi_{ij} = 1/5$ chance of being sampled represents $w_{ij} = 1/\pi_{ij} = 5$ similar students in finite population
  - Estimate of finite population mean (Horvitz-Thompson):

$$\widehat{\overline{y}}^{\mathrm{FP}} = \frac{1}{\sum_{ij} w_{ij}} \sum_{ij} w_{ij} y_{ij}$$

  - Similar for proportions, totals, etc.

Model-based inference for complex surveys

- Target of inference is superpopulation parameter β
- View finite population as simple random sample from superpopulation (or as realization from model)
- MLE $\widehat{\beta}^{\mathrm{FP}}$ using finite population treated as target (consistent for β)
- Design-based estimator of $\widehat{\beta}^{\mathrm{FP}}$ applied to complex survey data
  - Replace usual log likelihood by weighted log likelihood, giving **pseudo maximum likelihood estimator** (PMLE)
- If PMLE is consistent for $\widehat{\beta}^{\mathrm{FP}}$, then it is consistent for β

Multilevel modeling: Levels

- Levels of a multilevel model can correspond to stages of a multistage survey
  - Level-1: Elementary units i (stage 3), here students
  - Level-2: Units j sampled in previous stage (stage 2), here schools
  - Top-level: Units k sampled at stage 1 (primary sampling units), here areas
- However, not all levels used in the survey will be of substantive interest & there could be clustering not due to the survey design
- In PISA data, top level is geographical areas — details are undisclosed, so not represented as level in multilevel model

Two-level linear random intercept model

- Linear random intercept model for continuous y_{ij}:

$$y_{ij} = \beta_0 + \beta_1 x_{1ij} + \cdots + \beta_p x_{pij} + \zeta_j + \epsilon_{ij}$$

  - x_{1ij},. . .,x_{pij} are student-level and/or school-level covariates
  - β_0,. . .,β_p are regression coefficients
  - $\zeta_j \sim$ N$(0, \psi)$ are school-specific random intercepts, uncorrelated across schools and uncorrelated with covariates
  - $\epsilon_{ij} \sim$ N$(0, \theta)$ are student-specific residuals, uncorrelated across students and schools, uncorrelated with ζ_j and with covariates

Two-level logistic random intercept model

- Logistic random intercept model for dichotomous y_{ij}
  - As generalized linear model

$$\text{logit}[\text{Pr}(y_{ij} = 1|\mathbf{x}_{ij})] = \beta_0 + \beta_1 x_{1ij} + \cdots + \beta_p x_{pij} + \zeta_j$$

  - As latent response model

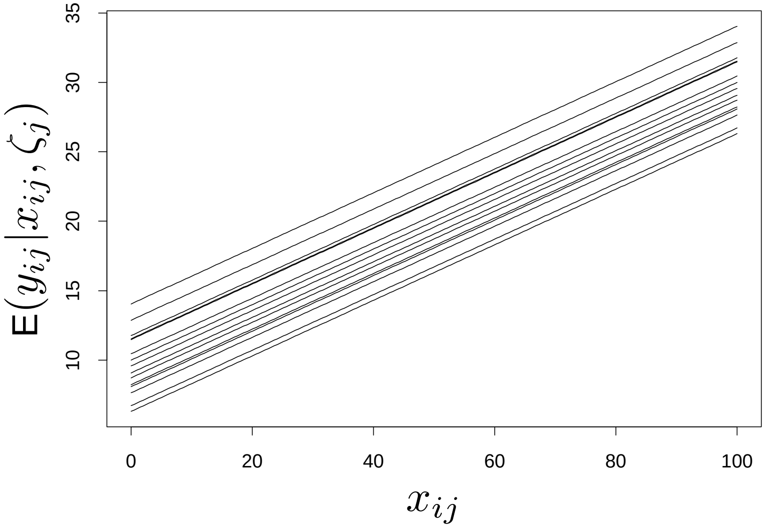
$$y^*_{ij} = \beta_0 + \beta_1 x_{1ij} + \cdots + \beta_p x_{pij} + \zeta_j + \epsilon_{ij}$$

$$y_{ij} = 1 \text{ if } y^*_{ij} > 0, \; y_{ij} = 0 \text{ if } y^*_{ij} \leq 0$$

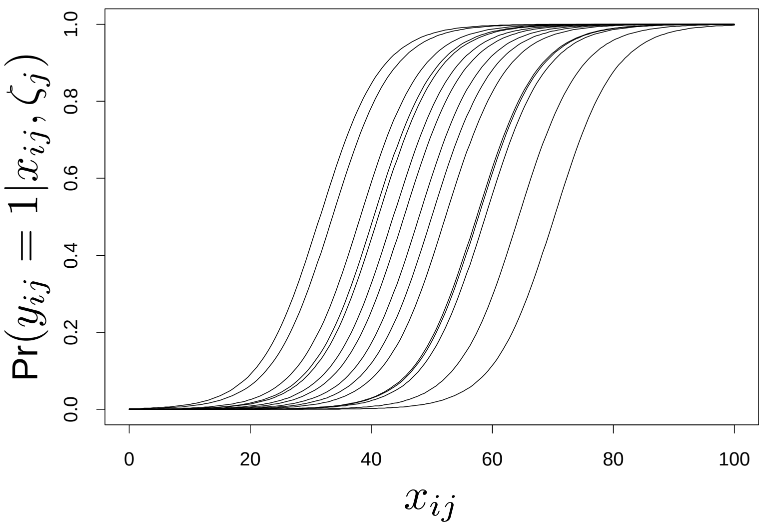
  - $\zeta_j \sim \text{N}(0, \psi)$ are school-specific random intercepts, uncorrelated across schools and uncorrelated with covariates
  - $\epsilon_{ij} \sim$ Logistic are student-specific residuals, uncorrelated across students and schools, uncorrelated with ζ_j and with covariates

Illustration of two-level linear and logistic random intercept model

$$\mathsf{E}(y_{ij}|x_{ij},\zeta_j) = \beta_0 + \beta_1 x_{ij} + \zeta_j$$



$$\mathsf{Pr}(y_{ij}=1|x_{ij},\zeta_j) = \frac{\exp(\beta_0+\beta_1 x_{ij}+\zeta_j)}{1+\exp(\beta_0+\beta_1 x_{ij}+\zeta_j)}$$



Pseudolikelihood

- Usual marginal log likelihood (without weights)

$$\log \prod_{j=1}^{n^{(2)}} \int \underbrace{\left\{ \prod_{i=1}^{n_j^{(1)}} f(y_{ij}|\zeta_j) \right\}}_{\Pr(\mathbf{y}_j|\zeta_j)} g(\zeta_j)\, \mathrm{d}\zeta_j = \sum_{j=1}^{n^{(2)}} \log \int \exp\left\{ \sum_{i=1}^{n_j^{(1)}} \log f(y_{ij}|\zeta_j) \right\} g(\zeta_j)\, \mathrm{d}\zeta_j$$

- Log pseudolikelihood (with weights)

$$\sum_{j=1}^{n^{(2)}} w_j \log \int \exp\left\{ \sum_{i=1}^{n_j^{(1)}} w_{i|j} \log f(y_{ij}|\zeta_j) \right\} g(\zeta_j)\, \mathrm{d}\zeta_j$$

- Note: need $w_j = 1/\pi_j, \quad w_{i|j} = 1/\pi_{i|j}$; cannot use $w_{ij} = w_{i|j} w_j$
- Evaluate using adaptive quadrature, maximize using Newton-Raphson [Rabe-Hesketh *et al.*, 2005] in `gllamm`

Standard errors, taking into account survey design

- Conventional “model-based” standard errors not appropriate with sampling weights
- **Sandwich estimator** of standard errors (Taylor linearization)

$$\text{Cov}(\widehat{\boldsymbol{\vartheta}}) \;=\; \mathcal{I}^{-1}\mathcal{J}\mathcal{I}^{-1}$$

  - $\mathcal{J}$: Expectation of outer product of gradients, approximated using PSU contributions to gradients
  - $\mathcal{I}$: Expected information, approximated by observed information (‘model-based’ standard errors obtained from $\mathcal{I}^{-1}$)
- Sandwich estimator accounts for
  - Stratification at stage 1
  - Clustering at levels ‘above’ highest level of multilevel model
- Implemented in `gllamm` with `cluster()` and `robust` options

Analysis of U.S. PISA 2000 data

- Two-level (students nested in schools) logistic random intercept model for reading proficiency (dichotomous)
- PSUs are areas, sampling weights $w_{i|j}$ for students and w_j for schools provided
- Predictors:
  - [Female]: Student is female (dummy)
  - [ISEI]: International socioeconomic index
  - [MnISEI]: School mean ISEI
  - [Highschool]/ [College]: Highest education level by either parent is highschool/college (dummies)
  - [English]: Test language (English) spoken at home (dummy)
  - [Oneforeign]: One parent is foreign born (dummy)
  - [Bothforeign]: Both parents are foreign born (dummy)

Data structure and gllamm syntax in Stata

- Data strucure

```
. list id_school wt2 wt1 mn_isei isei in 28/37, clean noobs

   id_school      wt2        wt1     mn_isei    isei
           2   105.82   .9855073    47.76471      30
           2   105.82   .9855073    47.76471      57
           2   105.82   .9855073    47.76471      50
           2   105.82   1.108695    47.76471      71
           2   105.82   .9855073    47.76471      29
           2   105.82   .9855073    47.76471      29
           3   296.95   .9677663          42      56
           3   296.95   .9677663          42      67
           3   296.95   .9677663          42      38
           3   296.95   .9677663          42      40
```

- gllamm syntax

```
gllamm pass_read female isei mn_isei high_school college
   english one_for both_for, i(id_school) cluster(wvarstr)
   link(logit) family(binom) pweight(wt) adapt
```

PISA 2000 estimates for multilevel regression model

	Unweighted Maximum likelihood		Weighted Pseudo maximum likelihood		
Parameter	Est	(SE)	Est	($\mathrm{SE_R}$)	($\mathrm{SE_R^{PSU}}$)
β_0: [Constant]	-6.034	(0.539)	-5.878	(0.955)	(0.738)
β_1: [Female]	0.555	(0.103)	0.622	(0.154)	(0.161)
β_2: [ISEI]	0.014	(0.003)	0.018	(0.005)	(0.004)
β_3: [MnISEI]	0.069	(0.001)	0.068	(0.016)	(0.018)
β_4: [Highschool]	0.400	(0.256)	0.103	(0.477)	(0.429)
β_5: [College]	0.721	(0.255)	0.453	(0.505)	(0.543)
β_6: [English]	0.695	(0.283)	0.625	(0.382)	(0.391)
β_7: [Oneforeign]	-0.020	(0.224)	-0.109	(0.274)	(0.225)
β_8: [Bothforeign]	0.099	(0.236)	-0.280	(0.326)	(0.292)
ψ	0.272	(0.086)	0.296	(0.124)	(0.115)

Problem with using weights in linear models

- Linear variance components model, constant cluster size $n_j^{(1)} = n^{(1)}$

$$y_{ij} \;=\; \beta_0 + \zeta_j + \epsilon_{ij}, \qquad \mathsf{Var}(\zeta_j) = \psi, \qquad \mathsf{Var}(\epsilon_{ij}) = \theta$$

- Assume sampling independent of ϵ_{ij}, $w_{i|j} = a > 1$ for all i, j
- Get biased estimate of ψ:
  - Weighted sum of squares due to clusters

$$\mathrm{SSC}^{\mathrm{w}} \;=\; \sum_j (\overline{y}_{.j} - \overline{y}_{..})^2 = \sum_j (\zeta_j - \overline{\zeta}_{.})^2 + \sum_j (\overline{\epsilon}^{\mathrm{w}}_{.j} - \overline{\epsilon}^{\mathrm{w}}_{..})^2 \;=\; \mathrm{SSC}$$

  - Expectation of $\mathrm{SSC}^{\mathrm{w}}$, same as expectation of unweighted SSC

$$\mathsf{E}(\mathrm{SSC}^{\mathrm{w}}) \;=\; (n^{(2)} - 1)\left[\psi + \frac{\theta}{n^{(1)}}\right]$$

  - Pseudo maximum likelihood estimator

$$\widehat{\psi}^{\mathrm{PML}} = \frac{\mathrm{SSC}^{\mathrm{w}}}{n^{(2)}} - \frac{\widehat{\theta}^{\mathrm{w}}}{an^{(1)}} \;>\; \widehat{\psi}^{\mathrm{ML}} = \frac{\mathrm{SSC}}{n^{(2)}} - \frac{\widehat{\theta}^{\mathrm{ML}}}{n^{(1)}}$$

Explanation for bias and anticipated results for logit/probit models

- Clusters appear bigger than they are (a times as big)
  - Between-cluster variability in $\bar{\epsilon}^{\mathrm{w}}_{\cdot j}$ greater than for clusters of size $an^{(1)}$
  - This extra between-cluster variability in $\bar{\epsilon}^{\mathrm{w}}_{\cdot j}$ is attributed to ψ
  - However, if sampling at level 1 stratified according to ϵ_{ij}, e.g.

$$\pi_{i|j} \approx \begin{cases} 0.25 & \text{if } \epsilon_{ij} > 0 \\ 0.75 & \text{if } \epsilon_{ij} \leq 0 \end{cases}$$

    variance of $\bar{\epsilon}^{\mathrm{w}}_{\cdot j}$ decreases, and upward bias of $\widehat{\psi}^{\mathrm{PML}}$ decreases
- Bias decreases as $n^{(1)}$ increases
- In logit/probit models, anticipate that $|\widehat{\beta}^{\mathrm{PML}}|$ increases when $\widehat{\psi}^{\mathrm{PML}}$ increases; therefore biased estimates of β

Solution: Scaling of weights?

- Scaling method 1 [Longford,1995, 1996; Pfeffermann *et al.*, 1998]

$$w^*_{i|j} = \frac{\sum_i w_{i|j}}{\sum_i w^2_{i|j}} w_{i|j} \quad \text{so that} \quad \sum_i w^*_{i|j} = \sum_i w^{*2}_{i|j}$$

  - In linear model example with sampling independent of ϵ_{ij}, no bias

```
egen sum_w = sum(w), by(id_school)
egen sum_wsq = sum(wˆ2), by(id_school)
generate wt1 = w*sum_w/sum_wsq
```

- Scaling method 2 [Pfeffermann *et al.*, 1998]

$$w^*_{i|j} = \frac{n^{(1)}_j}{\sum_i w_{i|j}} w_{i|j} \quad \text{so that} \quad \sum_i w^*_{i|j} = n^{(1)}_j$$

  - In line with intuition (clusters do not appear bigger than they are)

```
egen nj = count(w), by(id_school)
generate wt1 = w*nj/sum_w
```

Simulations

- Dichotomous random intercept logistic regression (500 clusters, N_j units per cluster in FP), with

$$y_{ij}^* \;=\; \underbrace{1}_{\beta_0} + \underbrace{1}_{\beta_1} x_{1j} + \underbrace{1}_{\beta_2} x_{2ij} + \zeta_j + \epsilon_{ij}, \quad \psi = 1$$

- Stage 1: Sample clusters with probabilities

$$\pi_j \;\approx\; \begin{cases} 0.25 & \text{if } |\zeta_j| > 1 \\ 0.75 & \text{if } |\zeta_j| \leq 1 \end{cases}$$

- Stage 2: Sample units with probabilities

$$\pi_{i|j} \;\approx\; \begin{cases} 0.25 & \text{if } \epsilon_{ij} > 0 \\ 0.75 & \text{if } \epsilon_{ij} \leq 0 \end{cases}$$

- Vary N_j from 5 to 100, 100 datasets per condition, 12-point adaptive quadrature

Results for $N_j = 5$

Parameter	True value	Unweighted ML	Weighted Pseudo maximum likelihood		
			Raw	Method 1	Method 2
		Model parameters: Conditional effects			
β_0	1	0.40	1.03	0.68	0.75
		(0.11)	(0.19)	(0.16)	(0.15)
β_1	1	1.08	1.19	0.96	0.98
		(0.18)	(0.32)	(0.26)	(0.26)
β_2	1	1.06	1.22	0.94	0.96
		(0.22)	(0.35)	(0.25)	(0.26)
$\sqrt{\psi}$	1	0.39	1.47	0.58	0.70
		(0.37)	(0.21)	(0.31)	(0.30)

Effect of level-1 stratification method ($N_j = 10$)

- (1) Strata based on sign of ϵ_{ij}
- (2) Strata based on sign of ξ_{ij}, Cor$(\epsilon_{ij}, \xi_{ij}) = 0.5$
- (3) Strata based on sign of ξ_{ij}, Cor$(\epsilon_{ij}, \xi_{ij}) = 0$

	True	Raw			Method 1		
Parameter	value	(1)	(2)	(3)	(1)	(2)	(3)
β_0	1	1.04	1.10	1.29	0.83	0.88	1.01
		(0.16)	(0.16)	(0.21)	(0.14)	(0.13)	(0.16)
β_1	1	1.06	1.11	1.26	0.91	0.92	0.99
		(0.23)	(0.26)	(0.30)	(0.20)	(0.23)	(0.25)
β_2	1	1.11	1.12	1.17	0.91	0.91	0.96
		(0.20)	(0.21)	(0.25)	(0.16)	(0.17)	(0.19)
$\sqrt{\psi}$	1	1.19	1.33	1.77	0.40	0.61	0.98
		(0.13)	(0.15)	(0.15)	(0.34)	(0.24)	(0.16)

Simulation results for pseudo maximum likelihood estimation

- Little bias for $\sqrt{\psi}$ when $N_j \geq 50$ (cluster sizes in sample $n_j^{(1)} \geq 25$)
- For smaller cluster sizes:
  - Raw level-1 weights produce positive bias for $\sqrt{\psi}$
  - Scaling methods 1 and 2 overcorrect positive bias for $\sqrt{\psi}$
    – apparently due to stratification based on sign of ϵ_{ij}
  - Inflation of $\boldsymbol{\beta}$ estimates whenever positive bias for $\sqrt{\psi}$
  - Good coverage using sandwich estimator (1000 simulations) for $N_j = 50$

Conclusions

- Pseudo maximum likelihood estimation allows for stratification, clustering, and weighting
- Three common methods for scaling level-1 weights: no scaling, scaling method 1, scaling method 2
- Inappropriate scaling can lead to biased estimates
  - If clusters are sufficiently large, little bias — similar results with all three scaling methods
  - If level-1 weights based on variables strongly associated with outcome, use no scaling
  - If level-1 weights based on variables not associated with outcome, use method 1
  - For intermediate situations, use method 2?

References

- Longford, N. T. (1995). *Models for Uncertainty in Educational Testing*. New York: Springer.
- Longford, N. T. (1996). Model-based variance estimation in surveys with stratified clustered designs. *Australian Journal of Statistics*, **38**, 333–352.
- Pfeffermann, D., Skinner, C. J., Holmes, D. J., Goldstein, H., & Rasbash, J. (1998). Weighting for unequal selection probabilities in multilevel models. *Journal of the Royal Statistical Society, Series B*, **60**, 23–40.

References: Our relevant work

- Rabe-Hesketh, S. & Skrondal, A. (2006). Multilevel modeling of complex survey data. *Journal of the Royal Statistical Society (Series A)* **169**, 805–827.
- Rabe-Hesketh, S., Skrondal, A. and Pickles, A. (2005). Maximum likelihood estimation of limited and discrete dependent variable models with nested random effects. *Journal of Econometrics* **128**, 301–323.
- Skrondal, A. & Rabe-Hesketh, S. (2004). *Generalized latent variable modeling: Multilevel, longitudinal and structural equation models*. Boca Raton, FL: Chapman & Hall/ CRC.
- `gllamm` and manual from `http://www.gllamm.org`