

Lecture 11

Incorporating complex survey design
information into MLMs

Lecture 11 Outline

- Stratified Survey
- Cluster Survey
- Define sampling weights
- Analysis of survey data
 - Descriptive inference
 - Design based approach
 - Analytic Inference
 - Design based approach
 - Model based approach
 - What do we do in MLMs?
 - Example of MLM: PISA 2000 US Data
 - My practical recommendations

Stratified Survey Design: An Example

- Program for International Student Assessment (PISA)
 - Conducted by the National Center for Educational Statistics
 - Coordinated by the Organization for Economic Development
 - System of international assessments that focus on 15-year-olds' capabilities in reading literacy, mathematics literacy, and science literacy.
 - 43 countries participate

PISA: US

- Data from 2000 US survey
 - We will look at this later as an example
- Each administration focuses on one of the three subject areas in depth but assesses each of the three subject areas
- <http://nces.ed.gov/surveys/pisa/index.asp>

PISA: Requirements for representative sample

- Represent the full population of 15-year-old students in each participating country and jurisdiction.
- Population is defined internationally as 15-year-olds attending both public and private schools in grades 7-12.
- Requires a minimum of 4,500 students from a minimum of 150 schools in each participating country and jurisdiction.
- Within schools, a sample of 35 students must be selected in an equal probability sample unless fewer than 35 students age 15 are available (in which case all students are selected).
- The school response rate target is 85 percent for all countries and jurisdictions.
 - A minimum participation rate of 65 percent of schools from the original sample of schools
 - A minimum participation rate of 80 percent of sampled students from schools within each country and jurisdiction

US Study Design

- *U.S. sampling frame*
 - Common Core of Data (CCD) listing of public schools supplemented with the Private School Universe Survey (PSS) listing of private schools. Close to 100 percent complete listing of schools
- *U.S. sampling design in 2009*
 - Stratified systematic sample
 - First define 8 strata defined by school type (public or private) and region (Northeast, Central, West, Southeast).
 - Students within strata are homogeneous
 - Schools are sampled from each strata, proportional to size
 - Second, students sampled at random from each school

Stratified Sample

- Partition the population into *more* homogeneous groups
 - States, ethnicities, education level, SES
- Sample units within each strata
- This design reduces sampling error of strata-specific rates since all groups of interest are included in the sample
- This design will decrease statistical variance, i.e. increases precision *for estimates of strata-specific rates*
- This design is more efficient than a simple random sample *to estimate strata-specific rates*
 - Need smaller total sample size when you stratify relative to using a simple random sample

Cluster Sample

- Divide the population of interest into geographic areas or clusters
 - Often times use Census tract as the cluster
- Take a sample of the clusters
- Measurements are taken on a sample within each selected cluster
- WHO 30 x 7 design
 - Randomly select 30 clusters from the list of potential clusters
 - Randomly select 7 interview locations within each cluster
- Persons within clusters are heterogeneous; ideal if cluster means are similar
- Not an efficient design
 - Need larger sample size relative to performing a simple random sample of the entire population
 - But this design is convenient!

Sampling weights

- Reflect the probability of the sampling unit (person, school, etc) of being selected for inclusion in the sample
- Typically presented as $w = 1 / Pr(selection)$
- In PISA 2009 design,
 - Assigned to each school in each strata
 - Roughly proportional to the size of the school
 - Students are selected at random from each school, so sampling weight is same for each student in selected school

Sampling Weights: Be Aware!

- Very common that the sampling weights contain additional adjustments!
 - Non-response: school-level, student-level
 - Imputation of item non-response within student assessments
 - Calibrate the sample to external control population
 - Weight trimming, to minimize impact of weights on estimation

Survey design produces correlated responses!

- Stratified random survey: PISA
 - Within strata (contextual variables for given types of people)
 - Within school correlation (contextual variables at the school level)
- Cluster survey
 - Within cluster (contextual variables based on geographic location)
 - Within person (genetic and temporal associations in individual responses over time)

Analysis of Survey Data

- Descriptive inference:
 - Most folks agree that the survey design is important to incorporate into descriptive statistics via survey weights
 - Known as “Design-based inference”
 - Variance estimates for descriptive statistics are then also adjusted for the design
- Analytic inference:
 - Hypothesis testing within regression model
 - Quite a bit of debate over the appropriate methods for incorporating the survey design into regression based analysis
 - Design-based inference vs. Model-based inference

To Weight or not To Weight?

That is the Question

- Question, Question, Question
- Bias - Variance Trade-off
- Beware highly variable weights
- In regression problems, regress on the weights too rather than weighting by them

Design-based descriptive inference

- Goal design-based inference:
 - estimate parameters associated with the target population incorporating features of the design
 - “impute” response for the non-sampled subjects
 - results generalize to target population
 - e.g. estimate the proportion of students that achieve reading proficiency in PISA
- Observed data and the sampling weights are the foundation of the design-based descriptive inference

Simple example

Assume 4 students from
Northeast public schools

Two schools

Design a study to estimate
total number with
proficiency in reading

One possible probability
sample: choose any pair of
persons with equal
probability

Label	School	Reading Proficiency
1	1	1
2	1	0
3	1	1
4	2	1
Known	Known	Unknown

Since we are choosing pairs out of 4 subjects, each selected subject will have probability of $2/4$ to be selected, so weight is 2 for each subject.

Simple Random Sample: n=2 out of 4 total

Persons selected	Measured values	Estimated total	Pr(sample)
1,2	1,0	$2*1+2*0=2$	1/6
1,3	1,1	$2*1+2*1=4$	1/6
1,4	1,1	$2*1+2*1=4$	1/6
2,3	0,1	$2*0+2*1=2$	1/6
2,4	0,1	$2*0+2*1=2$	1/6
3,4	1,1	$2*1+2*1=4$	1/6

Total = $1/6 (2+4+4+2+2+4) = 3$

Randomization Distribution

Stratified Sample: 1 from each school

Persons selected	Measured values	Estimated total	Pr(sample)
1,4	1,1	$3*1+1*1=4$	1/3
2,4	0,1	$3*0+1*1=1$	1/3
3,4	1,1	$3*1+1*1=4$	1/3

Total = $1/3 (4+1+4) = 3$

Randomization Distribution

Design-based estimate of total

- Horvitz-Thompson (unbiased and consistent) estimator for the population total
 - Foundation for all design-based analyses

$$\hat{T} = \sum_{i=1}^n \frac{1}{\pi_i} Y_i = \sum_{i=1}^n w_i Y_i$$

$$Y_i = \begin{cases} 0 & \text{not proficient} \\ 1 & \text{proficient} \end{cases}$$

π_i = probability that person i is selected in the sample

$w_i = \frac{1}{\pi_i}$ = sampling weight for person i

n = number of subject sampled

Design-based estimator of population mean

$$N = \sum_{i=1}^n \frac{1}{\pi_i} \text{ is known}$$

$$\hat{T} = \sum_{i=1}^n \frac{1}{\pi_i} Y_i$$

$$\bar{Y} = \hat{T} / N$$

Straight-forward to estimate variance of the mean for the population!

Design-based inference

- Most descriptive statistics of interest can be calculated based on totals/means
- Variance estimates:
 - Can get tricky!
 - Taylor series linearization
 - Well defined theory for all survey designs
 - Requires calculation of partial derivatives which may be hard
 - Replication methods
 - Jackknife methods: leave one out
 - Bootstrap or Balanced Repeated Replication: special case of jackknife when two clusters per strata
 - Easy to implement

Available Software

- SAS:
 - Proc's SURVEYSELECT, SURVEYMEANS, SURVEYFREQ
- R:
 - “survey” package
 - Then svymean, svytotal, svyby
- Stata:
 - svyset
 - Then svy: mean, proportion, ratio, total

Analysis of Survey Data

- Descriptive inference:
 - Most folks agree that the survey design is important to incorporate into descriptive statistics via survey weights
 - Known as “Design-based inference”
 - Variance estimates for descriptive statistics are then also adjusted for the design
- Analytic inference:
 - Hypothesis testing within regression model
 - Quite a bit of debate over the appropriate methods for incorporating the survey design into regression based analysis
 - Design-based inference vs. Model-based inference

Analytic Inference

- Goal is to specify a model for $f(Y|X, \pi)$ where $w = 1/\pi$ represent the sampling weights (provide information about design).
- When can we ignore the weights/design and use standard regression methods:

- Okay if design is non-informative

$$f(Y|X, \pi) = f(Y|X)$$

- Produce biased results if design is informative

$$f(Y|X, \pi) \neq f(Y|X)$$

What to do if design is informative?

- If the design is informative,
 - Design-based approach
 - Specify a model: $f(y_i|x_i)$
 - Then weight observations according to π_i
 - Weighted regression: Natural extension of descriptive inference based on randomization distribution
 - Model-based approach
 - Incorporate design features into the regression model itself
 - Model $f(y_i|x_i, \pi_i)$ directly
 - i.e. some function of the weights are included as a predictor, as are strata information, etc.
 - Likelihood based approach

Design-based approach

- Estimators are approximately unbiased and robust to model misspecification
- Inefficient; larger standard errors than model-based methods
- Hypothesis testing: limited to Wald type tests
- Model diagnostics: must think critically
- Most popular statistical software packages provide appropriate options

Model-based approach

- Likelihood based estimation
- Efficient; smallest standard errors if model is specified correctly
- Does not require specialized software
- Model diagnostics: we already know these
- Inference: likelihood ratio tests
- Results will be sensitive to model misspecification!
 - Missing key design elements

Multi-level models for survey data

Consider a simple MLM

- PISA data
- Ignore the strata for now
- Level 1 variable: student
- Level 2 variable: school

$$y_{ij} = \beta_0 + U_{0i} + \varepsilon_{ij}$$

$$U_{0i} \sim N(0, \tau^2)$$

$$\varepsilon_{ij} \sim N(0, \sigma^2)$$

and assume random effects are independent

i denotes the school, j denotes the student and y_{ij} is the reading proficiency score (continuous) for student j within school i

Recall estimation in mixed models

- Maximize the likelihood

$$L(\beta_0, \sigma^2, \tau^2 \mid Y) \quad \text{What we observe!}$$

$$= \prod_i \int \prod_j f(y_{ij} \mid U_{0i}, \beta_0, \sigma^2) f(U_{0i} \mid \tau^2) dU_{0i}$$

What we assume to be true

In the linear case, there is closed form solution

In non-linear case, there is no-closed form solution.

Requires special estimation procedures

Design-based mixed models?

- You could include key design features within the model
 - i.e. fixed effects of strata, we are already clustering by school
 - Ignore sampling weights if you think they are non-informative
- **Design-adjusted mixed models**
 - Extension of the design-based methods to mixed models
 - **Weighted likelihood approach**
 - **Pseudo-likelihood**

Sampling weights

- Sampling weights represent the probability of selection (plus additional corrections) and are measured at the various levels within the hierarchy

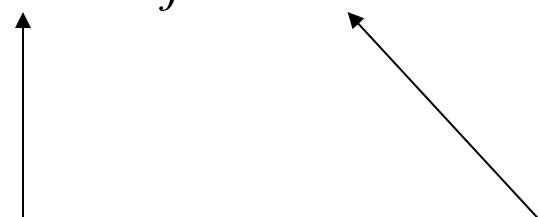
$$w_{ij} = \frac{1}{\pi_{ij}}, \pi_{ij} = \pi_i \times \pi_{j|i}$$

$$w_i = \frac{1}{\pi_i}$$

$$w_{j|i} = \frac{1}{\pi_{j|i}}$$

- In some cases, we may not have available the selection probability information at each level
 - In many publicly available datasets, we just get the overall weight

Pseudo-likelihood Approach

$$L(\beta, \sigma^2, \tau^2 \mid Y) = \prod_i w_i \int \prod_j w_{j|i} f(y_{ij} \mid x_{ij}, U_{0i}, \beta, \sigma^2) f(U_{0i} \mid \tau^2) dU_{0i}$$


School-level weight

Weight for student within school i

See Pfefferman, 1993, Feder, Nathan, Pfefferman 2000 and Rabe-Hesketh, Skrondal 2006

How do we handle the weights?

$$\text{Level 2 weight : } w_i = \frac{1}{\pi_i}$$

$$\text{Level 1 weight : } w_{j|i} = \frac{1}{\pi_{j|i}}$$

$$\text{So that : } w_{ij} = w_i w_{j|i}$$

How do we handle the weights?

NOTE: estimation within the mixed models framework MAY be sensitive to rescaling of the weights

This was NOT an issue in marginal models

Rescaled level-1 weights are recommended by
Pfeffermann et al (1998) and Skinner and Holmes (2003)

Several proposed scaled weights for instance :

$$w_{j|i}^r = \frac{w_{j|i}}{\overline{w}_{j|i}}$$

So that the average rescaled weights within school i is 1.

Available Statistical Software

- Stata: gllamm, mixed, **melogit**, **meqrlogit**
- HLM:
 - <http://www.ssicentral.com/hlm/index.html>
 - \$425
- M-plus:
 - <http://www.statmodel.com/>
 - \$600-\$1000

Special Note at this point

- Say that sampling of clusters was a simple random sample
 - i.e. each cluster had same probability of selection
 - Then we can ignore these weights in the MLM analysis
 - But if the units within the clusters are not a simple random sample, then the level 1 weight is important
- Say that the selection of units within a cluster are a simple random sample
 - i.e. each unit is selected with the same probability
 - Then we can ignore the level-1 weight
 - But if clusters are sampled proportional to size or any other non-equal probability method, then level-2 weights need to be considered.

Example: PISA 2000 US sample REVISED

- The goals of the analysis will be to:
 - quantify the variation in reading proficiency across U.S. schools
 - determine if the composition of the students within the schools explains the observed variation across schools (excluding SES)
 - quantify the effect of SES (via the contextual effect)
- Account for the survey design within the analysis and determine if the results are sensitive to inclusion of the design elements.

Data includes the following:

id:	school id (numeric)
pass_read:	dummy variable for being proficient in reading (1: proficient, 0: not)
female:	dummy variable for student being female
isei:	international socioeconomic index
high_school:	dummy variable for highest education level by either parent being high school
college:	dummy variable for highest education level by either parent being college
test_lang:	dummy variable for English being spoken at home (NOTE: the reading proficiency test is administered in English)
one_for:	dummy variable for one parent being foreign born
both_for:	dummy variable for both parents being foreign born
w_fstuwt:	student-level survey weight
wnrshbq:	school-level survey weight

Outcome: indicator for proficiency in reading

1. Random intercept for school and no covariates
 - Goal here is to estimate the variance of the random intercept for school
2. Compositional effects (composition of X values within cluster affects cluster-average outcome)
 - Add the school level summaries of gender, highest level of education among parents (high_school, college), foreign born parents (one_for, both_for) and English is primary language at home (test_lang)
 - Estimate the variance of the random intercept for school
3. Contextual effects of SES (cluster average X affects the individual outcomes)
 - Model 2 plus the within and between school SES variables
 - Estimate the contextual effect by taking the difference in the between effect and within effect.

Covariates*	Random Intercept Only	Random Intercept plus school characteristics	Random Intercept plus school characteristic plus SES
Proportion Female		0.98 (0.87, 1.09)	1.04 (0.94, 1.15)
Proportion of parents with High School education		1.14 (0.95, 1.36)	1.03 (0.87, 1.21)
Proportion of parents with College education		1.37 (1.16, 1.62)	1.04 (0.88, 1.22)
Proportion with one foreign born parent		1.02 (0.83, 1.26)	0.90 (0.75, 1.09)
Proportion with both foreign born parents		1.16 (0.98, 1.36)	1.07 (0.93, 1.25)
Proportion speaking English at home		1.36 (1.09, 1.69)	1.24 (1.01, 1.51)
Student SES (centered)			1.02 (1.01, 1.02)
Mean SES			1.08 (1.06, 1.11)
Contextual effect: SES			1.06 (1.04, 1.09)
Random intercept variance	0.82 (0.17)	0.44 (0.11)	0.23 (0.08)

- The covariates represent the school-level proportion of or mean characteristic unless otherwise noted. The odds ratios for the school-level proportion of the characteristic represent the relative odds per 10% difference in the characteristic.

Incorporate the Survey Weights

```
bys id: egen mean_schwt = mean(w_fstuwt)
gen wt1 = w_fstuwt / mean_schwt
rename wnrschbw wt2
```

* Random intercept - weighted

```
gllamm pass_read, i(id) family(binomial) adapt pweight(wt) from(a)
```

* Compositional model (ignoring SES) - weighted

```
gllamm pass_read mean_female mean_high_school mean_college ///
mean_one_for mean_both_for mean_test_lang, i(id) family(binomial) ///
adapt from(b) pweight(wt) eform
```

* Contextual effect of SES - weighted

```
gllamm pass_read isei_c mean_isei mean_female mean_high_school
mean_college ///
mean_one_for mean_both_for mean_test_lang, i(id) family(binomial) ///
adapt from(c) pweight(wt) eform
lincom mean_isei - isei_c, eform
```


Comparison of unweighted and weighted results

Covariates*	Unweighted Contextual Model	Weighted Contextual Model
Proportion Female	1.04 (0.94, 1.15)	1.23 (1.03, 1.46)
Proportion of parents with High School education	1.03 (0.87, 1.21)	1.03 (0.87, 1.21)
Proportion of parents with College education	1.04 (0.88, 1.22)	0.96 (0.76, 1.20)
Proportion with one foreign born parent	0.90 (0.75, 1.09)	1.01 (0.80, 1.28)
Proportion with both foreign born parents	1.07 (0.93, 1.25)	1.07 (0.84, 1.35)
Proportion speaking English at home	1.24 (1.01, 1.51)	1.17 (0.86, 1.59)
Student SES (centered)	1.02 (1.01, 1.02)	1.02 (1.01, 1.03)
Mean SES	1.08 (1.06, 1.11)	1.10 (1.05, 1.15)
Contextual effect: SES	1.06 (1.04, 1.09)	1.08 (1.03, 1.13)
Random intercept variance	0.23 (0.08)	0.27 (0.11)

How do you know the design is informative?

- Use your knowledge of the design and your research question and plan accordingly
- You could assess impact of accounting for the weights by:
 - Refit the model ignoring the sampling weights/clusters
 - Compare the model results; looking for substantial changes in estimated regression coefficients
 - Simple statistic to calculate is $|\hat{\beta}_w - \hat{\beta}_{nw}|/se(\hat{\beta}_{nw})$
 - If this is “large” then you know the weights are informative
 - We know the standard error estimates will generally increase when we add sampling weights