ANALYSIS OF A TWO-STAGE STRATIFIED CLUSTER SAMPLING DESIGN IN A NATIONAL SURVEY

Case Study: PMA2020 Ethiopia (Workie et al., 2017)

BY ARMIN MAHDAVI

STUDY OVERVIEW

Full title: A binary logistic regression model with complex sampling design of unmet need for family planning among all women aged (15-49) in Ethiopia.

- Population: Women aged 15–49 in Ethiopia
  - Sample size: n = 7,494
  - Outcome: Unmet need for family planning (binary)
    - 1= unmet need, 0=no unmet need.
    - All other background characteristics of the women were categorical
  - Objective: to estimate prevalence and factors associated with said unmet needs of family planning.
  - Specific parameter: population log-odds of unmet need among Ethiopian women aged 15-49.

SAMPLING DESIGN OVERVIEW

- Data Utilized: PMA2020 Round 4 Survey, a national conducted survey in Ethiopia, conducted in April 2016.
  - Performance Monitoring for Action, collaboration between Addis Ababa U and John Hopkins U.

- Two-stage stratified cluster sampling design
  - Stage 1: 221 Enumeration Areas (PSUs, clusters, or neighborhoods)
  - Stage 2: 35 households per EA (SRS within cluster)
  - Eligible women interviewed
  - Broad representation ensured but also introduced clustering effects via Stage 1.

Primary Sampling Unit = PSU

SAMPLING FRAME

- Central Statistical Agency's master sampling frame of enumeration areas (221 EAs selected).

STRATIFICATION

- PMA2020 population divided into strata
  - Improves precision
  - Ensures subgroup representation
  - Changes variance estimation structure

WHY SAMPLING DESIGN MATTERS

- Classical regression assumes IID sampling,
  - Independent (one observation's outcome isn't influencing another observation's outcome)
  - Equal selection probabilities
  - Simple Random Sampling (SRS)
  - Survey used stratification, clustering, multistage sampling

CLUSTERING AND INTRA-CLASS CORRELATION

- Women within same EA correlated
  - Design Effect: $Deff = 1 + (n-1)\rho$
  - Inflates variance relative to SRS
  - Ignoring clustering underestimates Standard errors!
    - $SE_{cluster} = \sqrt{Deff} \cdot SE_{SRS}$
  - Deff =1 same as …

Measures variance inflation when combined with cluster size via within-cluster similarity.

n= average cluster size (35)

ρ = intra-class correlation coefficient (ICC).

ρ = 1 → all group members are identical
ρ = 0 → none of the group members are identical (SRS)

- Example: ρ = 0.2, given the nature of these clusters,

And given n=35, Deff=2.79 so the clustering SE increases by a factor of 2.79

CLUSTERING AND INTRA-CLASS CORRELATION

- Effective sample size to consider,
- design effect "affects" effective sample size:
- $n_{eff} = \frac{n}{Deff}$
- With n=7,494, the "effective" sample size that can provide useful independent information is reduced by design effect value. This is because of ICC

ρ = intra-class correlation coefficient (ICC).

If women within clusters (same EA) are very similar the coefficient is higher → higher DE. This means a larger sample is now required to maintain same power of useful independent information.

IF SIMPLE RANDOM SAMPLING (SRS) WAS USED?

CONSIDERATION	EFFECT BY SRS	EFFECT BY TWO-STAGE CLUSTER
COST	HIGHER	LOWER
VARIANCE	LOWER	HIGHER
INDEPENDENCE	YES	NO
WEIGHTING	NOT NEEDED	NEEDED

WHY STANDARD LOGISTIC REGRESSION IS UNSUITABLE

Our logistic regression model is :

$$\text{logit}(\pi(\vec{X})) = \log\left(\frac{\pi(\vec{X})}{1-\pi(\vec{X})}\right) = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k$$

- logit(P(Y=1|X)) = β0 + β1X1 + ... + βpXp
  - Requires independence (cluster sampling violates)
  - Requires equal probability sampling
  - Estimator coefficients may remain consistent

COMPLEX OR SURVEY-ADJUSTED LOGISTIC REGRESSION (PSEUDO-LIKELIHOOD)

- Incorporates sampling weights
  - Accounts for clustering (PSU)
  - Accounts for stratification
  - Accounts for nonresponse adjustment

Base Weight: wi = 1 / πi

πi = selection probability

Corrects unequal probability sampling

provides population-representative estimates

Higher selection probability → less weight contribution and vice versa

- Variance estimated via first-order Taylor linearization at the PSU level.
  - Approximates sampling distribution of $\hat{\beta}$ to estimate its variance.
  - Accounts for clustering
- B-weighted information matrix (bread)
- M-cluster-level variation (meat)

Sandwich estimate of $\mathrm{cov}(\hat{\beta})$

$$\mathrm{co\hat{v}}(\hat{\beta}) = \hat{B}^{-1}\hat{M}\hat{B}^{-1}$$

$$\hat{B} = \sum_{i=1}^{N} \hat{D}_i \hat{V}_i^{-1} \hat{D}_i \quad \text{bread}$$

meat

$$\hat{M} = \sum_{i=1}^{N} \hat{D}_i \hat{V}_i^{-1} \{(Y_i - \hat{\mu}_i)(Y_i - \hat{\mu}_i)\} \hat{V}_i^{-1} \hat{D}_i$$

Consistent estimate of the true Covariance matrix of Y

HOW IS STANDARD ERROR AFFECTED? POINT ESTIMATES?

- All 7494 women interviewed for standard logistic regression assumes independent information. But with cluster sampling, women might likely be similar within same EAs. Design effect inflates variance.
- Less independent information available → variance of point estimates increases
  - Ignoring clustering underestimates Standard errors!
    - $SE_{cluster} = \sqrt{Deff} \cdot SE_{SRS}$
- Conditional mean model is unaffected.
- Point estimates themselves however, are often not changed.
    - Example of what that would look like:

Logistic Regression	$\hat{\beta}$	SE	p-value
Normal/naive	~same	lower	<0.05
Survey-adjusted	~same	higher	<0.05

SURVEY-WEIGHTED ESTIMATING EQUATION

- $\sum_{i=1}^{n} w_i X_i (Y_i - \mu_i) = 0$
- Where,
- $\mu_i = logit^{-1}(X_i^T \cdot \beta)$
- Survey logistic regression uses weighted score equation to solve for parameter estimates. This weight component reflects unequal selection probabilities mentioned earlier.

KEY RESULTS, VIA ODDS RATIOS

- Inference is design-adjusted
- Adjusted Odds Ratios Findings (of unmet family planning needs):
  - Urban residence ↓ odds
  - Young age ↑ odds
  - Currently married ↑ odds
  - Higher education ↓ odds
  - More births ↓ unmet need

The key determinants of unmet need FP in Ethiopia were residence, age, marital-status, education, household members, birth-events and number of under-5 children.

Government should take immediate steps to address the causes of high unmet need for family planning among women.

Table 5 Factors associated with unmet need for FP among aged (15-49) in Ethiopia, March to April, 2016

Variables		Total Unmet Need No	Yes n (%)	COR (95% CI)	AOR (95% CI)	Over all P-value
Place of Residence	Urban	1672	130(7.2)	0.33 (0.273, 0.399)*	0.717 (0.556, 0.924)*	<0.0001
	Rural	4605	1087(19.1)	1	1	
Age in groups	15-24	2766	313(10.2)	0.466 (0.397, 0.547)*	2.266 (1.644, 3.123)*	<0.0001
	25-34	1906	514(21.2)	1.112 (0.959, 1.288)	1.237 (0.989, 1.547)	
	35+	1605	390(19.5)	1	1	
Marital Status	Currently married	3549	1121(24.0)	19.5 (13.648, 27.863)*	8.077 (5.21, 12.523)*	<0.0001
	Living with man	87	25(22.3)	17.4 (9.857, 30.718)*	7.344 (3.654, 14.761*	
	Divorced/separated	488	28(5.4)	3.492 (2.077, 5.87)*	1.498 (0.825, 2.722)	
	Widow	189	12(6.0)	3.785 (1.900, 7.542)*	1.434 (0.663, 3.103)	
Never married	1962	32(1.6)	1	1		
highest level of school attended	Never	2418	701(22.5)	4.816 (2.182, 10.628)*	0.323 (0.186, 0.559)*	<0.0001
	Primary	2479	406(14.1)	2.724 (1.231, 6.026)*	0.417 (0.240, 0.726)*	
	Secondary	1076	93(8.0)	1.434 (0.635, 3.241)	0.499 (0.278, 0.896)*	
	Technical	191	10(5.0)	0.907 (0.332, 2.476)	0.460 (0.184, 1.150)	
	Higher	109	7(6.0)	1	1	

REPERCUSSIONS OF COMPLEX DESIGN OMISSION

- Underestimated standard errors
  - Inflated Type I error
  - Different conclusions
  - Potential policy misinterpretation

DESIGN STRENGTHS

- National representativeness
  - Cost-efficient implementation
  - Made region estimates possible
  - Design-consistent inference

LIMITATIONS

- Stratification variable not explicitly described in article
  - Perhaps due to PMA2020 protocol
- Specifics of weight construction and design effect value not provided
- Was FPC used? Not provided.

BROADER STATISTICAL IMPLICATIONS AND CONCLUSION

- Design-based vs Model-based inference
  - Survey logistic regression (inference-protected) vs Multilevel modeling (can model cluster variation)
  - Finite population inference
  - Public health policy implications

- Proper use of multistage stratified sampling
  - Correct survey-weighted logistic regression
  - Valid design-adjusted inference
    - Population of Ethiopian women in 2016 is fixed and known

CITATIONS

- Workie, D. L., Zike, D. T., Fenta, H. M., & Mekonnen, M. A. (2017). A binary logistic regression model with complex sampling design of unmet need for family planning among all women aged (15-49) in Ethiopia. *African Health Sciences*, *17*(3), 637–646. https://doi.org/10.4314/ahs.v17i3.6
- Week 6.ppt, STA 507, McClintock, S.
- Alimohamadi Y, Sepandi M. Considering the design effect in cluster sampling. J Cardiovasc Thorac Res. 2019;11(1):78. doi: 10.15171/jcvtr.2019.14. Epub 2019 Feb 17. PMID: 31024678; PMCID: PMC6477104.
- https://www.slideserve.com/irisa/estimation-in-marginal-models-gee-and-robust-estimation
- Dey D, Haque MS, Islam MM, Aishi UI, Shammy SS, Mayen MSA, Noor STA, Uddin MJ. The proper application of logistic regression model in complex survey data: a systematic review. BMC Med Res Methodol. 2025 Jan 22;25(1):15. doi: 10.1186/s12874-024-02454-5. PMID: 39844030; PMCID: PMC11752662.

QUESTION TIME