Discussion: Statistical Inference with Clustered Data in Survey Sampling

Michael P. Cohen*

Abstract
This is a summary of the discussion for the invited paper session "Statistical Inference with Clustered Data in Survey Sampling" at the 2016 Joint Statistical Meetings. The session was sponsored by the Survey Research Methods Section of the American Statistical Association and cosponsored by Statistics Without Borders. It was organized by Jae-kwang Kim and chaired by Zhengyuan Zhu, both from Iowa State University. The focus was on two-stage statistical survey models with clustering of primary sampling units.

Key Words: Statistical surveys, clusters, two-level models, correlation

1. Introduction

Session 101 at the 2016 Joint Statistical Meetings was an invited session on "Statistical Inference with Clustered Data in Survey Sampling" for which I was the discussant. Here I will summarize the pertinent points from the talks and give my comments and questions. I will do so in the order presented which was as follows:

1. Inference with Cluster Data Under Informative Sampling
   (Statistical inference using generalized linear mixed models under informative cluster sampling)
2. Bayesian Analysis for Cluster Sampling
3. H-Likelihood Method for Analyzing Clustered Survey Data

2. Inference with Cluster Data Under Informative Sampling

Jae-kwang Kim of Iowa State University was the presenter. His coauthors were Seunghwan Park and Youngjo Lee, both of Seoul National University. The basic model they treat has two levels with randomness at each of the levels.

Basic model:

$$y_{ij}|v_i \sim f_1(y_{ij}|\boldsymbol{x}_{ij}, v_i; \boldsymbol{\theta}_1), \quad j = 1, \ldots, M_i$$

unit or "student" level

$$v_i \sim f_2(v_i|\boldsymbol{z}_i; \boldsymbol{\theta}_2), \quad i = 1, \ldots, N$$

cluster or "school" level

They describe how such models are treated computationally when there is no informative sampling within clusters.

*American Institutes for Research, 1000 Thomas Jefferson Street NW, Washington DC 20007-3835

- With independent and identically distributed (i.i.d.) sampling, the EM algorithm finds the maximum likelihood estimator (MLE). It maximizes

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{N} \ell_i(\boldsymbol{\theta}).$$

- With sampling only at cluster level, the EM algorithm finds the pseudo-MLE. It maximizes

$$\ell_p(\boldsymbol{\theta}) = \sum_{i \in A^{(1)}} w_i \ell_i(\boldsymbol{\theta})$$

  where $A^{(1)}$ is the set of indices for the sampled clusters.

- But how do we use the EM algorithm with informative sampling within clusters?

Brief summary of presented solution:

1. "Pretend" the cluster-level random errors v_i are fixed but unknown parameters.
2. Estimate the v_i by pseudo-MLE weighted by $w_{j|i}$. Call these $\hat{v}_i$.
3. Now do the EM algorithm conditioning on $\hat{v}_i$. But we need to know the conditional distributions given $\hat{v}_i$ or an approximation.
4. The conditional distributions can be computed if we let

$$\hat{v}_i(\hat{\boldsymbol{\theta}}_1)|v_1 \sim N(v_i, V\{\hat{v}_i(\hat{\boldsymbol{\theta}}_1)\}).$$

5. Use the classical variance estimator for $V\{\hat{v}_i(\hat{\boldsymbol{\theta}}_1)\}$.

Discussion of this talk:

In response to this interesting presentation, I had some questions and a suggestion:

- Does this algorithm converge monotonically, at least approximately?
- Did the presenter ever encounter convergence problems?
- Has the presenter tried examples where the m_i are not all equal, $i = 1, \ldots, n$?
- The paper proposes a method to account for $\hat{\boldsymbol{\theta}}_1$ being an estimate in computing $V\{\hat{v}_i(\hat{\boldsymbol{\theta}}_1)\}$. Has this been tried on data?
- Of course, more simulations and real data examples would be welcome!

3. Bayesian Analysis for Cluster Sampling

Susanna Makela of Columbia University presented. Her coauthors were Yajuan Si of the University of Wisconsin–Madison and Andrew Gelman of Columbia University.

I will begin by giving some advantages of Bayesian approaches that are a blend of what was presented and my own views.

Advantages of Bayesian Approaches

- Sometimes we really do have prior information

- Results are not dependent on asymptotics (large sample theory)
- Principled and unified procedures
- Ironically, computational feasibility (they used to be hard to compute)
- In my opinion, the somewhat nonparametric Bayesian approaches should be less troubling to design-based people.

The presenter described the "existing Bayesian approach," that is, the one prior to the research of the presentation.

"Existing Bayesian Approach"

- Consider a single-stage probability proportional to size (PPS) sample of n units out of a population of N units with outcome y_i and size measure M_i, $i = 1, \ldots, n$.
- Assume the total population $T_M = \sum_{i=1}^{N} M_i$ is known.
- Factor likelihood for outcome y, size M, and inclusion indicator I.
- Use methods like splines, Bayesian bootstrap, or Dirichlet process priors to model the likelihood factors.
- See, e.g., Zangeneh, Keener, and Little (2011).

New in this Work

- Cluster size measures M_i and cluster population sizes N_i are only known for clusters i in the sample.
- Sometimes the assumption is made that $M_i = N_i$ for all clusters i.
- Introduces negative binomial model for cluster size measures M_i.

Negative Binomial Model

- Model cluster size M by $f(M|k, p) \sim NegBin(k, p)$.
- Then probability of observing in the sample a cluster of size

$$M_j \sim 1 + NegBin(k + 1, p)$$

  (Patil and Rao, 1978).
- We estimate $k + 1$ and p from sample. How?
- How do we determine if the negative binomial model fits the data?

Other Matters

- Also we must model outcomes y_i. In examples, y_i is normally distributed with mean depending on parameters, constant variance.
- I recommend considering models where $y_i > 0$ and variance increases with y_i. Common in establishment surveys.
- Lots to be explored!

4. H-Likelihood Method for Analyzing Clustered Survey Data

Donghwan Lee of the Ewha Womens University was the presenter with coauthor Youngjo Lee of Seoul National University. This presentation had some similarity to the first one, but differs in treating models having a multivariate error structure allowing correlated clusters. There is less emphasis on computational methods.

Population Model

- N clusters indexed by i.
- M_i units or elements in cluster i indexed by j.
- Two stage cluster sampling:

$$y_{ij}|v_i \sim f_1(y_{ij}|v_i;\theta_1)$$

$$\boldsymbol{v} \sim f_2(\boldsymbol{v};\theta_2).$$

- Often assume $\boldsymbol{v} \sim N(0,\Sigma)$ where Σ is a non-diagonal covariance matrix that may depend on θ_2.

Goals of Inference

1. Estimate $\boldsymbol{\theta} = (\theta_1, \theta_2)$.
2. Predict $\boldsymbol{v}$.

Problem to Solve

- If the sample design for clusters and units within a cluster is known, the problem can be solved by "H-likelihood method" (hierarchical likelihood).
- Suppose we only know $w_i = 1/\pi_i$ for clusters i and $w_{j|i} = 1/\pi_{j|i}$ for units j within cluster i.
- Basic idea (very simplified) is that the log-likelihood can be written as the sum of two terms:

$$h_w(\boldsymbol{\theta}, \boldsymbol{v}) = \sum_i w_i \sum_j w_{j|i} \log f_1(y_{ij}|v_i;\theta_1) + \log f_2(\boldsymbol{v};\theta_2).$$

Questions and Suggestions

- Real data examples would be welcome.
- I originally asserted: "Often clusters are selected one per stratum (or perhaps 2 or 3) so correlated clusters may be less common than suggested." I came to realize during the presentations that from a model-based viewpoint, correlations are the rule rather than the exception.
- Any comments on computing algorithms, convergence, run times, etc.?
- Software availability?

5. In closing…

I am especially indebted to the organizer, Professor Jae-kwang Kim, for organizing and for choosing me to be the discussant for this interesting session. I also thank the chair Professor Zhengyuan Zhu, the other participants, and the audience.

mpcohen@juno.com mcohen@air.org

REFERENCES

Patil, G. P., and Rao, C. R. (1978), "Weighted Distributions and Size-Biased Sampling with Applications to Wildlife Populations and Human Families," *Biometrics*, 34, 179–189.

Zangeneh, S. Z., Keener, R. W., and Little, R. J. A. (2011), "Bayesian Nonparametric Estimation of Finite Population Quantities in Absence of Design Information on Nonsampled Units," in *Proceedings of the Survey Research Methods Section*, Alexandria VA: American Statistical Association, pp. 3429–3440.