Bayesian inference for sample surveys

Roderick Little

Trivellore Raghunathan

Module 11: Hierarchical Models for Cluster Sample Designs

Models for Cluster Sample Design

- Hierarchical models (two-stage)

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

$$\alpha_i \sim N(0, \sigma_\alpha^2)$$

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

$$j = 1, 2, \cdots, K_i$$

$$i = 1, 2, \ldots, C$$

$$prior : \pi(\mu, \sigma_\alpha, \sigma_\varepsilon)$$

$$Data : \{y_{ij}, j = 1, 2, \ldots, k_i; i = 1, 2, \ldots, c\}$$

$$Draws:$$

$$(1)\ \mu, \sigma_\alpha, \sigma_\varepsilon, \alpha_i, i = 1, 2, \ldots, c$$

$$(2)\ y_{ij} \sim N(\mu + \alpha_i, \sigma_\varepsilon^2),$$

$$j = k_i + 1, k_i + 2, \ldots, K_i$$

$$i = 1, 2, \ldots, c$$

$$(3)\ \alpha_i \sim N(0, \sigma_\alpha^2), i = c + 1, \ldots, C$$

$$Y_{ij} \sim N(\mu + \alpha_i, \sigma_\varepsilon^2), j = 1, 2, \ldots, K_i$$

Implementation

- Use any standard Bayesian software package (Winbugs, Openbugs, STAN, JAGS, PROC MCMC, PROC MIXED etc) to obtain the draws in Step (1)
- Step (2) involves drawing normal random variables using the parameters from Step 1
- Step (3) Involves drawing normal random variables using the parameters from Step 1
- Once the population is filled-in compute the finite population quantity of interest

Incorporating other design features

- Include weights as covariates

$$Y_{ij} = \mu + \alpha_i + \beta f(w_i) + g(w_i)\varepsilon_{ij}$$

$$f() \text{ and } g() \textit{ known functions}$$

- Stratification
  - Analyze each stratum separately and fill-in the non-sampled values
  - If the number of clusters within a stratum is small then treat the stratum specific parameters as random effects

Multistage designs

- Typically more than 2-stages are involved in selecting the elements from the population
- Example:
  - Goal: A national probability sample of adults with representation from every State
    - Draw a sample of Counties from every State
    - Draw a sample of census tracts within the sampled counties
    - Draw a sample of block groups within the sampled tracts
    - Draw a sample of blocks within the sampled block groups
    - Draw a sample of households within the sample blocks
    - Draw an adult from the sampled households
- Often, for confidentiality reasons, only the first stage (counties) may be released.

Models for three stage design

- Nested Hierarchical models

$$Y_{ijk} = \mu + \alpha_i + \beta_{j(i)} + \varepsilon_{ijk}$$

$$i : Stage\ 1$$

$$j : Stage\ 2\ nested\ within\ Stage\ 1$$

$$k : Individuals$$

$$k = 1, 2, \cdots, N_{ij}$$

$$j = 1, 2, \ldots, S_i$$

$$i = 1, 2, \ldots, P$$

$$\alpha_i \sim N(0, \sigma_\alpha^2)$$

$$\beta_{j(i)} \sim N(0, \sigma_\beta^2)$$

$$\varepsilon_{ijk} \sim N(0, \sigma_\varepsilon^2)$$

$$prior : \pi(\mu, \sigma_\alpha, \sigma_\beta, \sigma_\varepsilon)$$

(1) Draw parameters
(2) Fill in non-sampled elements in the sampled and nonsampled first and second stage units

Binary Outcomes

- Mixed Effects Logistic Model

$$Y_{ijk} \sim Ber(\theta_{ijk})$$

$$\theta_{ijk} = \Pr(Y_{ijk} = 1)$$

$$logit(\theta_{ijk}) = \mu + \alpha_i + \beta_{j(i)}$$

$$\alpha_i \sim N(0, \sigma_\alpha^2)$$

$$\beta_{j(i)} \sim N(0, \sigma_\beta^2)$$

$$prior : \pi(\mu, \sigma_\alpha, \sigma_\beta)$$

Poisson Outcomes

- Count type variables

$$Y_{ijk} \sim Poisson(\theta_{ijk})$$

$$log(\theta_{ijk}) = \mu + \alpha_i + \beta_{j(i)}$$

$$\alpha_i \sim N(0, \sigma_\alpha^2)$$

$$\beta_{j(i)} \sim N(0, \sigma_\beta^2)$$

$$prior : \pi(\mu, \sigma_\alpha, \sigma_\beta)$$

Remarks

- All Bayesian analysis of survey data involves 3 step process:
  1. Generate draws of the parameters from their posterior distribution (this is a traditional Bayesian analysis step)
  2. Fill-in the non-sampled values conditional the draws of the parameters (just use the model, treating parameters as known)
  3. Compute the population quantity of interest
- Step 2 involves book keeping of whether the unit is sampled or not, and to use appropriate draws of the random effects

Remarks

- Assumed that cluster sizes for the sampled and non-sampled clusters are known
- This may not be true. The following approximation may be used
- For sampled clusters define $K_i = k_i / f$ where f is some small number , for example, 0.01 or 0.005
- For non-sampled clusters, bootstrap from the sampled cluster sizes $K_1, K_2, \ldots, K_c$