Bayesian Inference for Surveys

Rod Little and Trivellore Raghunathan

Approximations for Computations

Two Stages

- Model for prediction
  - Joint distribution for the fine population values
    - Exchangeable on the index used to label units in the population
    - Introduce parameters, conditional on which , the units are independent
    - Prior on the parameters induces a joint distribution
- Analysis
  - Summary of the population values
  - Analytical models fitted to the entire population (linear regression). These also can be viewed as summary of the population values
- Model for prediction can lead to approximation of the summary of the population values

Examples of Approximation

- Stratified population

$$N_h, h = 1, 2, \cdots, H$$

$$Y_{ih}, i = 1, 2, \cdots, N_h$$

$$\bar{Y} = \sum_h \sum_i Y_{ih} / N, N = \sum_h N_h$$

- Prediction or Population Model

$$\prod_h P(Y_{i1,} Y_{i2}, \cdots, Y_{iN_h})$$

- Exchangeable within-stratum and independence across-stratum

$$\prod_h \left[\int \left(\prod_{i=1}^{N_h} f(Y_{ih} \mid \theta_h) \right) \pi(\theta_h) d\theta_h \right]$$

- Estimand

$$Q(Y) = \Pr(Y \geq c)$$

Analysis

- Conceptually, the straightforward approach is to generate several draws of nonsampled values, computed the proportion of the sampled and drawn values exceeding the constant c.
- Approximation
  - Prediction Model

$$Y_{ih} \mid \mu_h, \sigma_h^2 \sim iid\ N(\mu_h, \sigma_h^2)$$

$$\pi(\mu_h, \sigma_h) \propto \sigma_h^{-1}$$

- Model based approximation

$$\Pr(Y \geq c \mid \mu_h, \sigma_h, h = 1, 2, \cdots, H)$$

$$= \sum_h W_h \{1 - \Phi((c - \mu_h) / \sigma_h)\}$$

$$W_h = N_h / N$$

$$N_h \gg n_h$$

Draws

- A standard Bayesian analysis from each stratum to obtain draws of

$$(\mu_h, \sigma_h) \mid y_{ih}, i = 1, 2, \cdots, n_h$$

- Compute the approximation given on the previous slide to obtain approximate draws of

$$\Pr(Y \geq c)$$

- Though assumed normality, the approach works for any other parametric distribution with a known functional form
  - t-distribution
  - Chi-square or gamma
  - Beta
- Transformation to normality is another possible approach

$$\Pr(Y \geq c) \Rightarrow \Pr(g(Y) \geq g(c))$$

$$g(y) \sim Normal$$

Imputation Based Approaches

- Suppose that any of the standard parametric forms don't fit the data well and none of the transformation works well
- Tukey introduced a general class of models that can accommodate a wide variety of distributions

$$Y = \mu + \sigma \times Z \times \frac{\exp(gZ) - 1}{gZ} \times \exp(hZ^2 / 2)$$

- Parameter

$$\mu : Location$$

$$\sigma : Scale$$

$$g : Skewness$$

$$h : Kurtosis(tails)$$

It is very difficult to write the density function of Y but it is easy to draw from.

Estimation of the parameters is easy using the percentiles

Estimation

- Location parameter: Sample Median

$$p = \Pr(Y \leq Q_p)$$

$$A_p = \frac{Q_{1-p} - Q_{0.5}}{Q_{0.5} - Q_p} = \exp(-gZ_p)$$

$$B_p = \frac{g(Q_{1-p} - Q_{0.5})}{\exp(-gZ_p - 1)} = \sigma \exp(hZ_p^2 / 2)$$

Estimation (contd.)

- Choose various values of p and obtain sample percentiles

$$Regress\ \ \log(A_p) \text{ on } -Z_p \text{ to estimate } g$$

$$Regress\ \ \log(B_p) \text{ on } Z_p^2 \text{ to estimate } \sigma \text{ and } h$$

Imputation approach implemented in IVEware (Version 0.3 to be soon released)

- Draw an Approximate Bayesian bootstrap sample (Draw a bootstrap sample and again draw a bootstrap sample from this bootstrap sample)
- Estimate the percentiles and estimate the parameters
- Impute the unobserved values by drawing values of the standard normal random variables and using the estimated parameters from the previous step
- Stratum-specific imputations

Implementation

- For each stratum create the missing values for the unobserved subjects
- Use the “gh” command in IVEware and “by Stratum” command
- Create multiple imputations
- Compute the population quantity of interest

BBDESIGN

- IVEware (version 0.3) also implements a more general non-parametric approach for drawing the unobserved values in the population using Bootstrap and Polya urn model (that we discussed previously)
- This approach may be more preferable if good parametric models could not be found to fit the data well