Complex Survey Variance and Design Effects in R

using the Rstan and Survey packages

Matt Williams[1] Terrance Savitsky[2]

[1]National Center for Science and Engineering Statistics
National Science Foundation
mrwillia@nsf.gov

[2]Office of Survey Methods Research
Bureau of Labor Statistics
Savitsky.Terrance@bls.gov

GASP

Sept 23, 2019

Population Inference from Complex Survey Samples

- Goal: perform inference about a finite population generated from an unknown model, P_{θ_0}.
- Data: from under a complex sampling design distribution, P_ν
  - Probabilities of inclusion π_i are often associated with the variable of interest (purposefully)
  - Sampling designs are "informative": the balance of information in the sample $\neq$ balance in the population.
- Biased Estimation: estimate P_{θ_0} without accounting for P_ν.
  - Use inverse probability weights $w_i = 1/\pi_i$ to mitigate bias.
- Incorrect Uncertainty Quantification:
  - Failure to account for dependence induced by P_ν leads to standard errors and confidence intervals that are the wrong size.

Variance Estimation

- The de-facto approach:
  - approximate sampling independence of the primary sampling units (Heeringa et al. 2010).
  - within-cluster dependence treated as nuisance
- Two common methods:
  - Taylor linearization and replication based methods.
  - A variety of implementations are available (Binder 1996, Rao et al. 1992).

Taylor Linearization

Let y_{ij}, X_{ij}, and w_{ij} be the observed data for individual i in cluster j of the sample. Assume the parameter θ is a vector of dimension d with population model value θ_0.

1. Approximate an estimate $\hat{\theta}$, or a 'residual' $(\hat{\theta} - \theta_0)$, as a weighted sum: $\hat{\theta} \approx \sum_{i,j} w_{ij} z_{ij}(\theta)$ where z_{ij} is a function evaluated at the current values of y_{ij}, X_{ij}, and $\hat{\theta}$.
2. Compute the weighted components for each cluster (e.g., primary sampling units (PSUs)): $\hat{\theta}_j = \sum_i w_{ij} z_{ij}(\theta)$.
3. Compute the variance between clusters: $\widehat{Var(\hat{\theta})} = \frac{1}{J-d} \sum_{j=1}^{J} (\hat{\theta} - \hat{\theta}_j)(\hat{\theta} - \hat{\theta}_j)^T$
4. For stratified designs, compute $\hat{\theta}_s$ and $\widehat{Var(\hat{\theta}_s)}$ within strata and sum $\widehat{Var(\hat{\theta})} = \sum_s \widehat{Var(\hat{\theta}_s)}$.

Replication

Let y_{ij}, X_{ij}, and w_{ij} be the observed data for individual i in cluster j of the sample. Assume the parameter θ is a vector of dimension d with population model value θ_0.

1. Through randomization (bootstrap), leave-one-out (jackknife), or orthogonal contrasts (balanced repeated replicates), create a set of K replicate weights $(w_i)_k$ for all $i \in S$ and for every $k = 1, \ldots, K$.
2. Each set of weights has a modified value (usually 0) for a subset of clusters, and typically has a weight adjustment to the other clusters to compensate: $\sum_{i \in S}(w_i)_k = \sum_{i \in S} w_i$ for every k.
3. Estimate $\hat{\theta}_k$ for each replicate $k \in 1, \ldots, K$.
4. Compute the variance between replicates: $\widehat{Var(\hat{\theta})} = \frac{1}{K-d} \sum_{k=1}^{K} (\hat{\theta} - \hat{\theta}_k)(\hat{\theta} - \hat{\theta}_k)^T$.
5. For stratified designs, generate replicates such that each strata is represented in every replicate.

Challenges

There are two notable trade-offs associated with these methods:

- Taylor linearization: value $\hat{\theta}$ computed once then used in a plug in for $z_i(\theta)$.
  - Replication methods: estimate $\hat{\theta}_k$ computed K times.
  - Sizable differences in computational effort
- Replication methods: no derivatives are needed.
  - Taylor linearization: requires the calculation of a gradient to derive the analytical form of the first order approximation $z_i(\theta)$.
  - This poses significant analytical challenges for all but the simplest models.

Some Improvements

- Abstraction of Derivatives (less analytic work for Taylor Linearization)
  - Recent advances in algorithmic differentiation (Margossian 2018), allows us to specify the model as a log density but only treat the gradient in the abstract without specifying it analytically.
  - Implemented in Stan and Rstan (Carpenter 2015, Stan Development Team 2016)
- Hybrid Approach or Taylor Linearization for replicate designs (less computation for Replication approaches)
  - Survey package (Lumley 2016) to calculate replication variance of gradient
  - Use plug in for θ, only estimate once

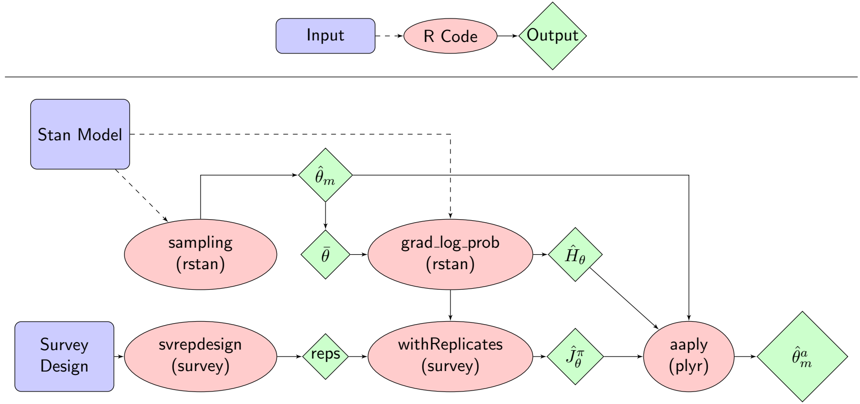
Example: Design Effect for Survey-Weighted Bayes

- Williams & Savitsky (2018): https://arxiv.org/abs/1807.11796
- Pseudo posterior $\propto$ Pseudo Likelihood $\times$ Prior

$$\pi^{\pi}\left(\boldsymbol{\lambda}|\mathbf{y},\tilde{\mathbf{w}}\right) \propto \left[\prod_{i=1}^{n} \pi\left(y_i|\boldsymbol{\lambda}\right)^{\tilde{w}_i}\right] \pi\left(\boldsymbol{\lambda}\right)$$

- Variances Differ:
  - Weighted MLE: $H_{\theta_0}^{-1} J_{\theta_0}^{\pi} H_{\theta_0}^{-1}$ (Robust)
  - Weighted Posterior: $H_{\theta_0}^{-1}$ (Model-Based)
- Adjust for Design Effect: $R_2^{-1} R_1$
  - $\hat{\theta}_m \equiv$ sample pseudo posterior for $m = 1, \ldots, M$ draws with mean $\bar{\theta}$
  - $\hat{\theta}_m^a = \left(\hat{\theta}_m - \bar{\theta}\right) R_2^{-1} R_1 + \bar{\theta}$
  - where $R_1' R_1 = H_{\theta_0}^{-1} J_{\theta_0}^{\pi} H_{\theta_0}^{-1}$
  - $R_2' R_2 = H_{\theta_0}^{-1}$

R Code Schematic



References I

Binder, D. A. (1996), 'Linearization methods for single phase and two-phase samples: a cookbook approach', *Survey Methodology* **22**, 17–22.

Carpenter, B. (2015), 'Stan: A probabilistic programming language', *Journal of Statistical Software* .

Heeringa, S. G., West, B. T. & Berglund, P. A. (2010), *Applied Survey Data Analysis*, Chapman and Hall/CRC.

Lumley, T. (2016), 'survey: analysis of complex survey samples'. R package version 3.32.

Margossian, C. C. (2018), 'A review of automatic differentiation and its efficient implementation', *CoRR* **abs/1811.05031**.
URL: *http://arxiv.org/abs/1811.05031*

Rao, J. N. K., Wu, C. F. J. & Yue, K. (1992), 'Some recent work on resampling methods for complex surveys', *Survey Methodology* **18**, 209–217.

Stan Development Team (2016), 'RStan: the R interface to Stan'. R package version 2.14.1.
URL: *http://mc-stan.org/*

Williams, M. R. & Savitsky, T. D. (2018), 'Bayesian uncertainty estimation under complex sampling', *arXiv preprint arXiv:1807.11796* .