Combining Probability and Nonprobability Samples Under Unknown Overlaps

Terrance D. Savitsky [1] Matthew R. Williams [2]
Julie Gershunskaya [3] Vladislav Beresovsky [4]
Nels G. Johnson [5]

[1] U.S. Bureau of Labor Statistics (Office of Survey Methods Research)

[2] RTI International (Division for Statistical and Data Sciences)

[3] U.S. Bureau of Labor Statistics (OEUS Methods Division),

[4] National Center for Health Statistics

[5] USDA Forest Service

FCSM

October, 2022

Outline

Combine convenience sample with reference sample

- Improve estimation efficiency with convenience sample
  - Non-probability sample inexpensive and easily accessible
  - Often has a lot more units than reference probability sample
- Treat convenience sample as from latent random sampling mechanism:
  - Estimate latent inclusion probabilities, $\pi_c(\mathbf{x}_i)$
  - Use overlap of predictor values $(\mathbf{x}_{ci}, \mathbf{x}_{ri})$ and known reference sample $\pi_r(\mathbf{x}_i)$
  - Reference and convenience samples may overlap units
- Exclude convenience units that inflate estimator variance
  - Remove convenience units very different from reference
  - $\mathbf{x}_{ci}$ values very different from $\mathbf{x}_{ri}$

Outline

Motivation

Methods

Bayesian Hierarchical Model

Simulation Performance Study

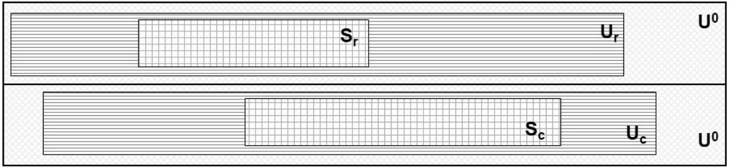
Terminology

- S_c and S_r observed convenience and reference samples
- Population frames U_c and U_r
- Target population U^0, such that $U_c \subseteq U^0$ and $U_r \subseteq U^0$.
- Known coverage probabilities of U^0 by frames U_c and U_r
  - $p_c(\mathbf{x}_i) = P\left\{i \in U_c | i \in U^0, \mathbf{x}_i\right\}$
    $p_r(\mathbf{x}_i) = P\left\{i \in U_r | i \in U^0, \mathbf{x}_i\right\}$
- inclusion probabilities into S_c and S_r
  - $\pi_c(\mathbf{x}_i) = P\{i \in S_c | i \in U_c, \mathbf{x}_i\}$
    $\pi_r(\mathbf{x}_i) = P\{i \in S_r | i \in U_r, \mathbf{x}_i\}$
- Consider combined sample, $S = S_c + S_r$.
- Indicator $z_i = 1$ when $i \in S_c$, and $z_i = 0$ when $i \in S_r$
- $\pi_z(i) = P\{i \in S_c \mid i \in S, \mathbf{x}_i\} \to$ propensity scores

Proposition: The following relationship holds:

$$\pi_z(\mathbf{x}_i) = \frac{\pi_c(\mathbf{x}_i)\, p_c(\mathbf{x}_i)}{\pi_c(\mathbf{x}_i)\, p_c(\mathbf{x}_i) + \pi_r(\mathbf{x}_i)\, p_r(\mathbf{x}_i)}.$$

Proof: two copies of U^0: $U = U^0 + U^0$.



$$P\{i \in S_c | i \in U, \mathbf{x}_i\} = P\{i \in S_c | i \in U_c, i \in U^0, \mathbf{x}_i\} P\{i \in U_c | i \in U^0, \mathbf{x}_i\} P\{i \in U^0 | i \in U\}$$
$$= \frac{1}{2} \pi_c(\mathbf{x}_i)\, p_c(\mathbf{x}_i)$$

Similarly, $P\{i \in S_r | i \in U, \mathbf{x}_i\} = \frac{1}{2}\pi_r(\mathbf{x}_i)\, p_r(\mathbf{x}_i)$.

Hence, for units in $S = S_r + S_c$, we have

$$\begin{aligned} P\left\{i \in S | i \in U, \mathbf{x}_i\right\} &= P\left\{i \in S_c | i \in U, \mathbf{x}_i\right\} + P\left\{i \in S_r | i \in U, \mathbf{x}_i\right\} \\ &= \frac{1}{2}\pi_c\left(\mathbf{x}_i\right) p_c\left(\mathbf{x}_i\right) + \frac{1}{2}\pi_r\left(\mathbf{x}_i\right) p_r\left(\mathbf{x}_i\right). \end{aligned}$$

By definition of conditional probability,

$$P\left\{i \in S_c | i \in S, i \in U, \mathbf{x}_i\right\} = \frac{P\left\{i \in S_c | i \in U, \mathbf{x}_i\right\}}{P\left\{i \in S | i \in U, \mathbf{x}_i\right\}}$$

Exact Likelihood Method when $U_c = U_r$

- Same Population Frame for each sampling arm
- $p_c(\mathbf{x}_i) = P(i \in U_c \mid i \in U^0) = p_r(\mathbf{x}_i)$

$$\pi_z(\mathbf{x}_i) = \frac{\pi_c(\mathbf{x}_i)}{\pi_c(\mathbf{x}_i) + \pi_r(\mathbf{x}_i)}$$

- Produces exact likelihood for observed data $z_i \sim$ Bernoulli$(\pi_z(\mathbf{x}_i))$, which allows to implicitly estimate parameters of $\pi_c(\mathbf{x}_i, \beta)$
- Elliot, 2009 derived the same formula assuming no-overlap between samples
- S_c and S_r may be overlapping

Outline

Joint model for $[(z_i), (\pi_{ri})_{i \in S_r}]$

1. Parameterize our model using $\pi_{\ell i} = P\{i \in S_\ell \mid i \in U_\ell, \mathbf{x}_i\}$.
   - Unit $i \in 1, \ldots, (n = n_r + n_c)$
   - Sampling arm $\ell \in (r, c)$
   - Estimate $(\pi_{\ell i})$ for all units for both $\ell = r$ and $\ell = c$
2. logit$(\pi_{\ell i}) = \mu_{x,\ell i} = \mathbf{x}_i \boldsymbol{\gamma}_{x,\ell} + \sum_{k=1}^{K} g(x_{ki}) \boldsymbol{\beta}_{\ell k}$
   - B-spline basis for *each* predictor where $C \times 1$, $g(x_{ki})$, with $C =$ knots + spline degrees - 1
   - Autoregressive smoothing of the $C \times 1$, $\boldsymbol{\beta}_{\ell k}$
   - Sparsity over K predictors with $\beta_{\ell k c} \sim \mathcal{N}(\beta_{\ell k c-1}, \kappa_{\ell k} \tau_\ell)$
3. Joint likelihood for $[(z_i), (\pi_{ri})_{i \in S_c}]$
   - $z_i \mid \pi_{zi} \overset{\text{ind}}{\sim}$ Bernoulli(π_{zi})
   - logit$(\pi_{ri}) \overset{\text{ind}}{\sim} \mathcal{N}(\mu_{x,ri}, \phi)$ only for units $i \in S_r$

Outline

Motivation

Methods

Bayesian Hierarchical Model

Simulation Performance Study

Compare Exact and Pseudo Likelihood Methods

- Exact Likelihood Methods (Bayesian Implementation)
  - Two-arm option:
    $(S_c, S_r) : \pi_z(\mathbf{x}_i) = \pi_c(\mathbf{x}_i) / (\pi_c(\mathbf{x}_i) + \pi_r(\mathbf{x}_i))$
  - One-arm option: $(S_c, U) \rightarrow \pi_r(\mathbf{x}_i) = 1$: $\pi_z(\mathbf{x}_i) = \frac{\pi_c(\mathbf{x}_i)}{\pi_c(\mathbf{x}_i)+1}$
  - One-arm gold standard since know whole population of X.
- Pseudo Likelihood Methods (Bayesian Implementation)
  - Competitors define likelihood on population indicator
  - Approximate on observed sample using weights $\propto 1/\pi_r(\mathbf{x}_i)$
    - Chen, P. Li, and Wu (2020)(LCW) specify Bernoulli $(\pi_c(\mathbf{x}_i))$ for pop
    - Wang, Valliant, and Y. Li (2021) (WVL) specify Bernoulli $(\pi_z(\mathbf{x}_i))$ for pop - same as One-arm

Data Generation Process

- We generate $M = 30$ distinct populations of size $N = 4000$.
  - Let X have $K = 5$ predictors (one continuous)
  - Outcome y_i has a lognormal distribution $\log(y_i) \sim \mathcal{N}(\mathbf{x}_i\beta, 2)$.
- We chose a large sampling fractions to explore the full range of $\pi_c \in [0, 1]$ (establishment surveys).
  - Select reference sample of $n_r = 400$ using PPS sampling: $s_{r_i} = \log(\exp(\mathbf{x}_i\beta) + 1)$
  - Select two convenience samples of $n_c \approx 800$ using Poisson sampling: $\pi_{c_i} = \text{logit}^{-1}(\mathbf{x}_i\beta_c + \text{offset})$
  - We control 'high' and 'low' overlap by varying β_c compared to the reference sample (next slide)

High and Low Overlap of X_r and X_c Datasets

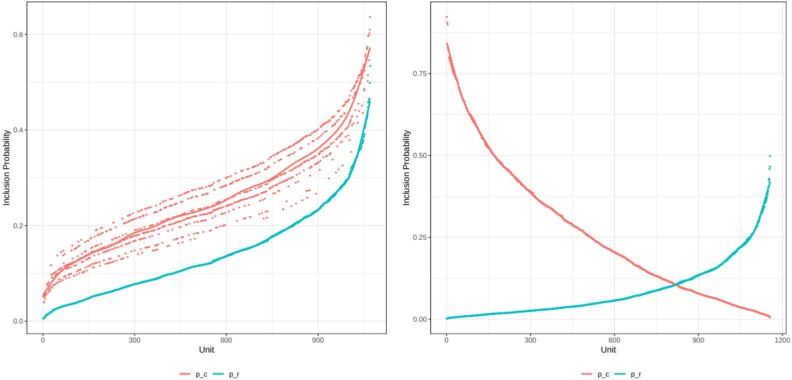


Figure: π_c versus π_r **LHS** high overlap and **RHS** low overlap.

Higher Percent of Pooled Sample in High Overlap

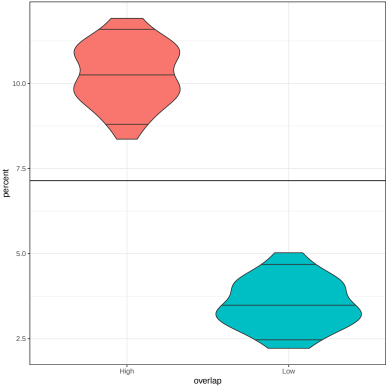


Figure: Distributions over 30 population and sample realizations.

Two-arm Method is More Efficient

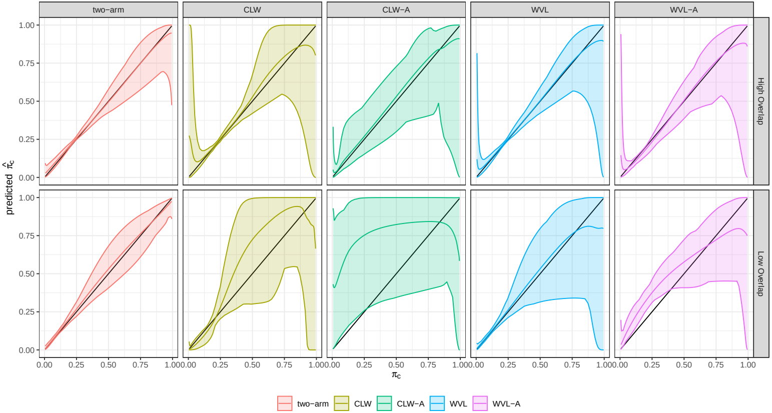


Figure: Avg and 95% frequentist quantiles for posterior mean of π_c .

Coverage degrades for pseudo likelihood options

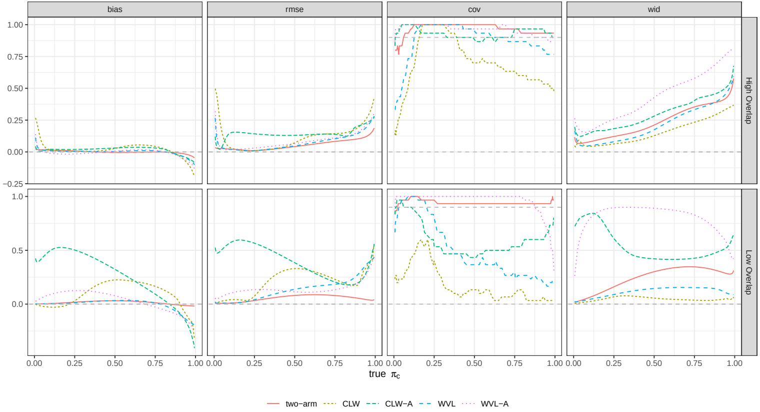


Figure: Pointwise coverage comparisons of 90% credibility intervals in 3rd column

Application to Estimation of Government Employment

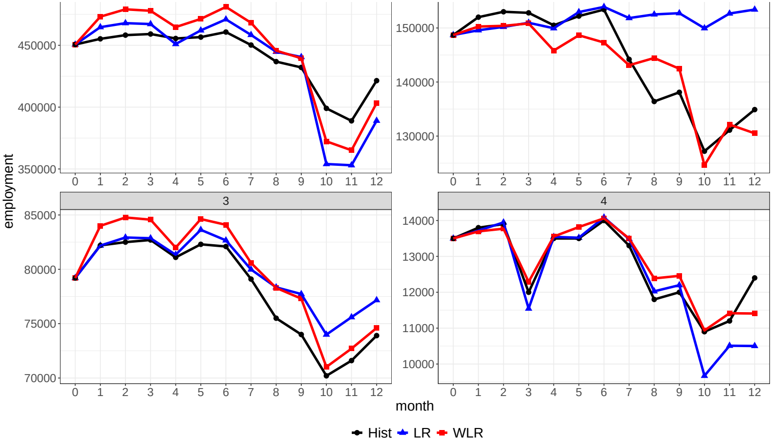
- Estimate pseudo weights for quota sample of government employment.
- Use census instrument as reference sample; we set $\pi_{ri} = 1$ for all units
- We observe: $z = 1$ for units in the quota sample and $z = 0$ for units in the census.
- Quota sample units are a subset of census.
- Estimate $\pi_c(\mathbf{x}_i)$ of inclusion into the quota sample, where $\mathbf{x}_i$ is employment level of unit i
- Produce employment estimates for Metropolitan Statistical Areas (MSAs).

Pseudo Weighted Link Relative Estimator (WLR)

$$\hat{Y}_{d,12} = Y_{d,0} \prod_{\tau=1}^{12} \hat{R}_{d,\tau}.$$

- Starting level, $Y_{d,0}$, available from census at end of year
- Monthly ratio estimates $\hat{R}_{d,\tau}$ are obtained using a link relative (LR) estimator
$$\hat{R}_{d,\tau}^{LR} = \sum_{i \in s_{d,\tau}} y_{i,\tau} / \sum_{i \in s_{d,\tau}} y_{i,\tau-1}$$
- We fear LR induces bias by use of unweighted $(y_{i,\tau-1}, y_{i,\tau})$.
- So, use a weighted LR estimator.
$$\hat{R}_{d,\tau}^{WLR} = \sum_{i \in s_{d,\tau}} w_i y_{i,\tau} / \sum_{i \in s_{d,\tau}} w_i y_{i,\tau-1}$$

Estimations for Selected MSAs



Future Work

- Joint estimation of (π_c, y).
- Create efficient survey estimator for domains.
- Incorporates full uncertainty quantification.

References I

Chen, Yilin, Pengfei Li, and Changbao Wu (2020). "Doubly Robust Inference With Nonprobability Survey Samples". In: *Journal of the American Statistical Association* 115.532, pp. 2011–2021.

Elliot, Michael R. (2009). "Combining Data from Probability and Non-Probability Samples Using Pseudo-Weights". In: *Survey Practice* 2 (6), pp. 813–845.

Wang, L., R. Valliant, and Y. Li (2021). "Adjusted logistic propensity weighting methods for population inference using nonprobability volunteer-based epidemiologic cohorts". In: *Stat Med.* 40.4, pp. 5237–5250.