A look at propensity-based methods for combining probability and non-probability sample data

Based on works of Savitsky, Gershunskaya, Beresovsky, Johnson, and others

Matthew R. Williams

RTI International (mrwilliams@rti.org)

Outline

Motivation

- Provide some insights on the topic of combining survey sample and convenience data
- Invite some discussion - there is a lot of work in this area and this presentation is a partial representation

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 process
  - Estimate pseudo inclusion probabilities, $\pi_c(\mathbf{x}_i)$
  - Then combine with reference sample (under $\pi_r(\mathbf{x}_i)$)
- Use convenience samples as augmented sample
  - Propagate weight estimation uncertainty into final analysis
- Reference and convenience samples may have duplication

Population Inference from Informative Samples

- Goal: perform inference about a finite population generated from an unknown model, $\mathbb{P}_{\theta_0}(\mathbf{y})$.
- Data: from under a complex sampling design distribution, $\mathbb{P}_{\nu}(\boldsymbol{\delta})$
  - Probabilities of inclusion $\pi_i = Pr(\delta_i = 1|\mathbf{y})$ 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 $\mathbb{P}_{\theta_0}(\mathbf{y})$ without accounting for $\mathbb{P}_{\nu}(\boldsymbol{\delta})$.
  - Use inverse probability weights $w_i = 1/\pi_i$ to mitigate bias.
- Incorrect Uncertainty Quantification:
  - Failure to account for dependence induced by $\mathbb{P}_{\nu}(\boldsymbol{\delta})$ leads to standard errors and confidence intervals that are the wrong size.

Outline

The Ecology Problem

- Use-Availablity (Johnson, Williams, and Riordan, 2021):
  - Convenience sample of plant species locations from herbarium collections. (Semi-purposeful or haphazard).
  - Habitat data (weather, soil conditions, etc) for an entire region
  - How can we predict suitability/presence of plant species for unvisited areas?
- Case-Control Studies (Lancaster and Imbens, 1996)
  - Want to match positive cases to controls (null cases) controlling for similar predictors (weak causality)
  - Controls are actually the entire population -don't know if any cases are present: Contaminated Controls.
- These are both the same problem!

The Ecology Problem

- Want to model the probability of a case $P(Y = 1|X)$
- Only observe Z - indicator of each source: $Z = 1$ for convenience cases, $Z = 0$ for the entire population.
- We need a mapping from $P(Z = 1|X)$ to $P(Y = 1|X)$

$$\pi_z(x_i) = \pi_y(\mathbf{x}_i)/(\pi_y(\mathbf{x}_i) + 1) \tag{1}$$

- Instead of doing a logistic regression (generalized linear) model $\pi_y(\mathbf{x}_i)$ we have a more irregular non-linear regression model for $\pi_z(x_i)$.
- Modelling $\pi_z(x_i)$ as a logistic (linear) regression: ranks might be preserved, but point estimates are biased. Ranks not preserved across multiple outcomes (species).
- Software is now available - Stan and BRMS (Burkner, 2017)

Outline

The Survey Sampling Problem

- We have a representative survey sample, optimized for key measures and sup-populations. May have insufficient sample size for sub-populations and additional measures of interest (rare responses).
- Much potential for augmenting with other data sources (convenience).

The Survey Sampling Problem

- Different approaches to combining based on available information
  - Propensity Approaches*: Assume have record level data for both sources
  - Weight Calibration: Use population totals
  - Record linkage: Assume have matched individuals across sources.
  - Mass Imputation: Use model fit on one source - port over to the other source.
- Different options for estimation after 'combining':
  - *Inverse-weighted estimation*
  - Doubly-Robust estimation
  - Model-based estimation

Application: Large Scale Volunteer Data (UK Biobank)

- UK Biobank - 500,000 participants but errors for sample coverage and low participation rates (5%)
- Correcting for volunteer bias in (Genomewide Association Studies) GWAS uncovers novel genetic variants and increases heritability estimates (Alten et al., 2022)
  - Used UK Census microdata as reference sample to estimate pseudo-probability weights.
- Addressing selection bias in the UK Biobank neurological imaging cohort (Bradley and Nichols, 2022)
  - Used Health Survey for England as a reference sample
  - Compared a variety of methods (calibration, logistic regression, BART) for estimating pseudo-probability weights
- Prevalance of 'Phecodes' across UKB, All of US, and Michigan Genomics Iniviative (Salvatore et al., 2024)

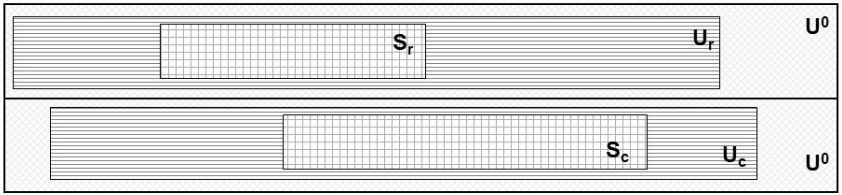
The Survey Sampling Problem

- We have selection probabilities for reference sample $P(r = 1|X)$
- Want to model the probability of being selected into the convenience sample $P(c = 1|X)$ (to use for inverse weighting)
- Only observe Z - indicator of each source: $Z = 1$ for convenience cases, $Z = 0$ for the reference sample.
- We need a mapping from $P(Z = 1|X)$ to $P(Y = 1|X)$

$$\pi_z(x_i) = \pi_c(\mathbf{x}_i)/(\pi_c(\mathbf{x}_i) + \pi_r(\mathbf{x}_i)) \tag{2}$$

- First proposed by (Elliott, 2009), our team relaxed and simplified the justification (Savitsky et al., 2023). More on next slide.

The Survey Sampling Problem



- Make two copies of the population U^0
- Have a sampling frame U_r and U_c
- Take samples from each frame S_r and S_c
- Overlapping units are sampled in each but not uniquely identified. The stacked population $U = U^0 + U^0$ is used to get the probabilities right.
- For simplicity of discussion, assume the frame and population are the same $U^0 = U_r = U_c$*

The Survey Sampling Problem

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

- Many applications directly model the propensity $\pi_z(x_i)$ as a logistic (linear) regression. Similar to our ecology applications, this leads to bias estimation of $\pi_c(x_i)$ - the intended target.
- If reference sample $\pi_r(x_i)$ are available for those in the convenience sample, we proceed as a non-linear regression (Beresovsky). (Stratified designs)
- If we don't know $\pi_r(x_i)$ for convenience units.
  - Pseudo-likelihood*: Wang, Valliant, and Y. Li (2021) and Chen, P. Li, and Wu (2020)
  - Joint-modelling: Savitsky et al. (2023)

* Some effort going to pseudo-posterior.

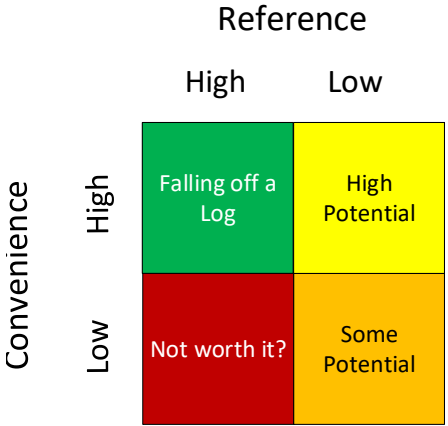
Compare Exact and Pseudo Likelihood Methods

- Exact* Likelihood Methods
  - One-arm option: $(S_c, U) \rightarrow \pi_r(\mathbf{x}_i) = 1$: $\pi_z(\mathbf{x}_i) = \frac{\pi_c(\mathbf{x}_i)}{\pi_c(\mathbf{x}_i)+1}$
  - 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 gold standard since know whole population of X.
- Pseudo Likelihood Methods
  - Chen, P. Li, and Wu (2020) specify Bernoulli $(\pi_c(\mathbf{x}_i))$ for pop.
  - Approximate on observed sample using weights $\propto 1/\pi_r(\mathbf{x}_i)$
  - Wang, Valliant, and Y. Li (2021) specify Bernoulli $(\pi_z(\mathbf{x}_i))$ for pop - same as One-arm
  - But use the reference sample to approximate population with weights $\propto 1/\pi_r(\mathbf{x}_i)$

Outline

Drivers for Success (Difficulty)

- Informativeness of two samples (efficiency from pooling data)
  - Both - High chance of success (two efficient samples!!!)
  - Neither - Significant improvement in estimation (still not great?)
  - Ref Inform but Con Not - Might not be worth it.
  - Con Inform but Ref Not - Significant improvement (might be good overall)
- Remember - this varies by outcome!



Drivers for Success (Difficulty)

- ‘Overlap’ of two samples (efficiency of estimating propensities)
  - High overlap - covariate distributions are very similar - several methods lead to good estimation of propensities
  - Low overlap - covariate distributions are very different - most methods struggle with estimating propensities. The ‘best’ one requires additional models or knowledge of π_r.

High and Low Overlap of X_r and X_c Datasets

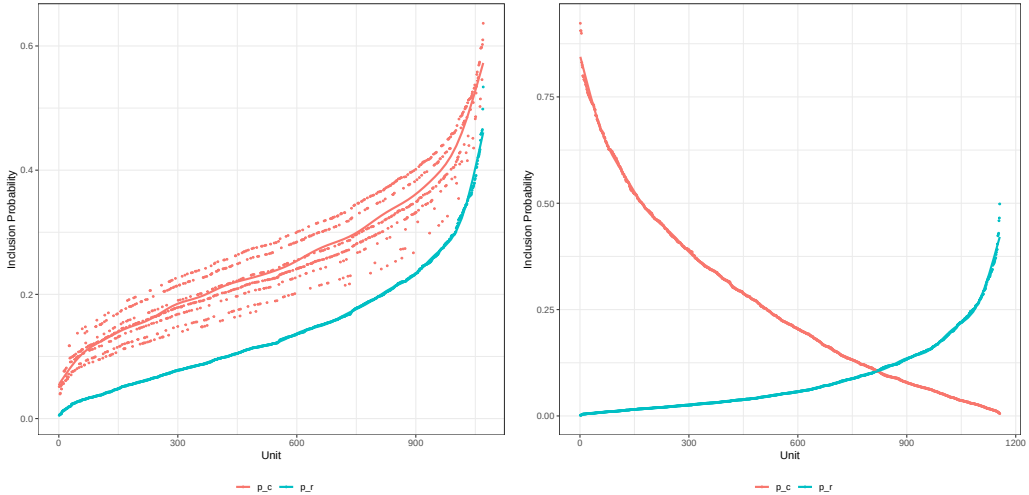


Figure: π_c versus π_r **LHS** high overlap and **RHS** low overlap.

High and Low Overlap of X_r and X_c Datasets

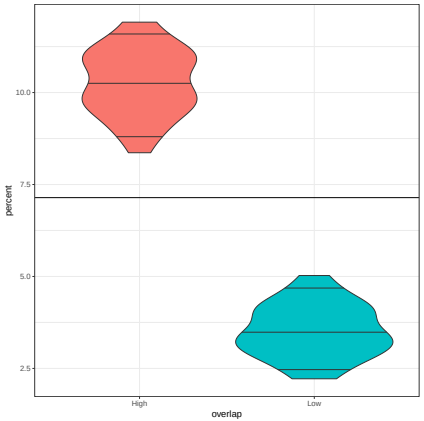
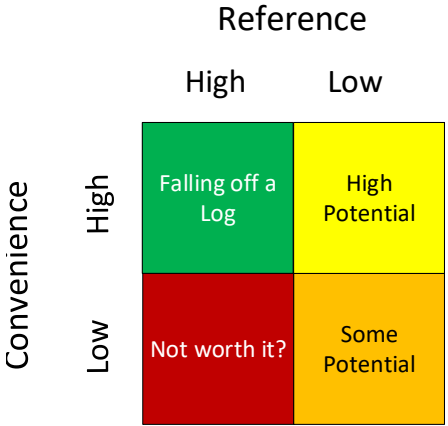


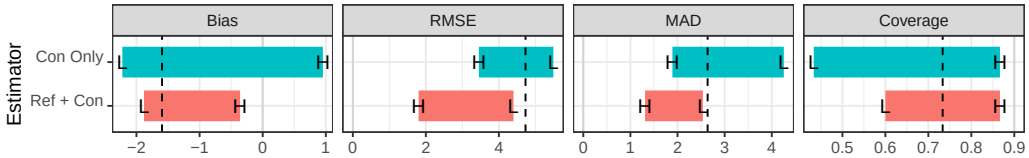
Figure: Percent of pooled sample present in both reference and convenience samples by type of convenience sample (High and Low). Expected percent for two independent simple random samples (solid horizontal line).

Drivers for Success (Difficulty)

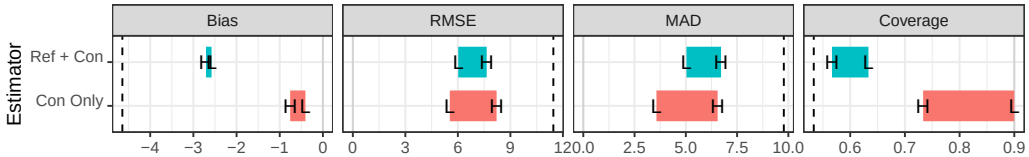
- Overlap and Informativeness are Related
  - High Overlap - likely both samples are either informative or not (Both Informative and High Overlap - 'easiest' case)
  - Low Overlap - likely only one sample is informative (Con Informative but Low overlap with Uninformative Ref - challenging to estimate but high potential for improvement)



Informative Reference Sample By Varying Overlap



Anti-Informative Reference Sample By Varying Overlap



Recap

- Don't model the propensity π_z as a logistic (linear) regression. Use a nonlinear model approach based on stacked populations. Use as flexible model as you can (trees, splines, etc).
- If reference and convenience samples have similar data distributions, either the pseudo or full approaches both help (assuming model for π_c is correct). Full approaches (with additional model assumptions) have potential to be more efficient.
- If reference and and convenience samples have low overlap (and low sample sizes), modelling for π_r (full approach) is needed. The final estimates (incorporating all uncertainty) are much better than simply using the reference sample if the reference is not very informative or useful for that outcome.
- We can't avoid modelling. We need to be careful.

References I

- Alten, Sjoerd van et al. (2022). “Correcting for volunteer bias in GWAS uncovers novel genetic variants and increases heritability estimates”. In: *medRxiv*.
- Bradley, Valerie and Thomas E. Nichols (2022). “Addressing selection bias in the UK Biobank neurological imaging cohort”. In: *medRxiv*.
- Burkner, Paul-Christian (2017). “brms: An R Package for Bayesian Multilevel Models Using Stan”. In: *Journal of Statistical Software* 80.1, pp. 1–28.
- 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.
- Elliott, Michael R. (2009). “Combining Data from Probability and Non-Probability Samples Using Pseudo-Weights”. In: *Survey Practice* 2 (6), pp. 813–845.
- Johnson, Nels G., Matthew R. Williams, and Erin C. Riordan (2021). “Generalized nonlinear models can solve the prediction problem for data from species-stratified use-availability designs”. In: *Diversity and Distributions* 27.11, pp. 2077–2092.

References II

- Lancaster, Tony and Guido Imbens (1996). “Case-control studies with contaminated controls”. In: *Journal of Econometrics* 71.1-2, pp. 145–160.
- Salvatore, Maxwell et al. (May 2024). “To weight or not to weight? The effect of selection bias in 3 large electronic health record-linked biobanks and recommendations for practice”. In: *Journal of the American Medical Informatics Association* 31.7, pp. 1479–1492.
- Savitsky, Terrance D. et al. (Dec. 2023). “Methods for combining probability and nonprobability samples under unknown overlaps”. English (US). In: *Statistics in Transition New Series* 24.5.
- 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.