Integrating Differnet Data Sources for Total Estimation with Unknown Population Size

Zhaoce (Charlie) Liu

ITSEW2019
Department of Statistical Science
Southern Methodist University

zhaocel@smu.edu

June 12, 2019

Overview

- Motivating Example
- Review of Current Methods
- Propensity Score Adjustment (PSA)
- Adaptive Propensity Score Adjustment (APSA)
- Simulation Study
- Questions

Motivating Example

The National Marine Fisheries Service (NMFS) estimates total fish caught by recreational anglers

- Marine Recreational Information Program (MRIP)
- Gulf of Mexico: AL, FL, LA, MS, TX
- Many species: Red Snapper, King Mackerel, White Grunt...
- Fishing season, bag limits

Motivating Example: Data Sources

- Dockside Intercept Survey: S_2
  - Catch per Unit Effort (CPUE): Average catch per species per trip
  - Probability proportional to size Design (PPS)
  - PSU: Dock's Location $\times$ Times $\times$ Days
  - High quality, small sample size, expensive
- Electronic Reporting Sample: S_1
  - Captains can volunteer to participate
  - Non-probability sample
  - Contains similar information as the Dockside Intercept Sample, including the response variable (y)
  - Low quality, large sample size, low cost
- Overlap between the two samples

Motivating Example: Data Sources Visualization

- Goal: Estimate the total fish caught with unknown population size

Review of Current Methods

- Liu et al. (2017): Ratio Estimators
- Use the self-reported sample as auxiliary information
- $\hat{t}_{yp} = \frac{n_1}{\hat{p}_1}\widehat{\bar{y}} = n_1 \frac{\sum_{i \in s_2} w_i y_i}{\sum_{i \in s_1 \cap s_2} w_i} = n_1 \frac{\hat{t}_y}{\hat{n}_1}, \quad \text{ratio } B_p = \frac{t_y}{n_1}$
- $\hat{t}_{yc} = t_{y^*} \frac{\sum_{i \in s_2} w_i y_i}{\sum_{i \in s_1 \cap s_2} w_i y_i^*} = t_{y^*} \frac{\hat{t}_y}{\hat{t}_{y^*}}, \quad \text{ratio } B_c = \frac{t_y}{t_{y^*}}$
- $\hat{t}_{MR} = (1 - w)\hat{t}_{yp} + w\hat{t}_{yc}$

Propensity Score Adjustment (PSA)

- Involove the non-probability sample into estimation directly
  - Major issue: Selection bias of the non-probability sample
  - Lee and Valliant, 2009; Elliott et al., 2017; Kim and Wang, 2018; etc
    - Combine the nonprobability sample with a probability sample
    - Create pseudo-weights for the non-probability sample
    - Variable of interest only available in the non-probability sample
  - Robbins, M. W.,2017
    - Overlap exists between the two samples
    - Joint weighting and disjoint weighting
- But...
  - How well are the samples integrated?
  - How much selection bias can be adjusted by PSA?

Adaptive Propensity Score Adjustment (APSA)

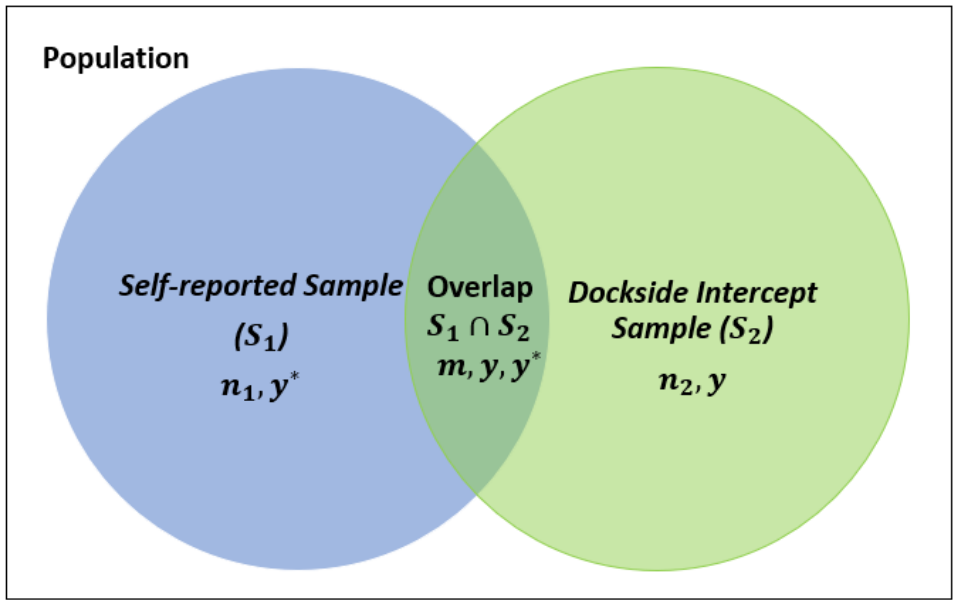
- We propose a new propensity-score-based weighting approach
  - Take advantage of the response variable from the probability sample
  - Monitor the sample integration process
  - Detect the non-representative part of the non-probability sample

Adaptive Propensity Score Adjustment (APSA)

5-Step Adaptive Propensity Score Adjustment Method

- Step 1. Calculate the propensity score for every unit in the combined sample
- Step 2. Sort the estimated propensity scores from smallest to largest and segment the sample into 10 subgroups by decile points
- Step 3. Within each subgroup, compare the conditional distributions of response variable between units from both samples, we use K-S test here
- Step 4. Identify subgroup with significant test p-value, discard units from the non-probability sample but keep the units from the probability sample from that subgroup
- Step 5. Re-calculate propensity score for the remaining data and conduct the general PSA procedure

General Framework



- Notation:
  - y: variable of interest
  - $\mathbf{x}$: $q \times 1$ vector of covariates available in both S_1 and S_2
  - $S_1^* = S_1 \cap S_2^c$
  - $\alpha_i = P\left(i \in S_2 | \mathbf{x}_i\right), \ i \in S_2$
  - Propensity score: $\gamma_i = P(i \in S_1^* | S_1^* \cup S_2), \ i \in S_1^* \cup S_2$

General Framework

- To estimate the pesudo-inclusion probability for the non-probability sample
  - $\beta_i = P(i \in S_1 | \mathbf{x}_i),\ i \in S_1$
  - $q_i = P(i \in S_1^* | \mathbf{x}_i)$
  - $p_i = P\left(i \in S_1^* \cup S_2 | \mathbf{x}_i\right),\ i \in S_1^* \cup S_2$
- $\hat{\beta}_i = \frac{\hat{\alpha}_i \hat{\gamma}_i}{(1-\hat{\alpha}_i)(1-\hat{\gamma}_i)}),\ i \in S_1$
- $\hat{q}_i = \frac{\hat{\alpha}_i \hat{\gamma}_i}{1-\hat{\gamma}_i}, i \in S_1^*$
- $\hat{p}_i = \begin{cases} \hat{\alpha}_i + \hat{\beta}_i - \hat{\alpha}_i \hat{\beta}_i = \frac{\hat{\alpha}}{1-\hat{\gamma}_i} & i \in S_1^* \\ \alpha_i + \hat{\beta}_i - \alpha_i \hat{\beta}_i = \frac{\alpha_i}{1-\hat{\gamma}_i} & i \in S_2 \end{cases}$

General Framework

- Joint Weighting:
  - The samples are combined to be representive of the population
  - $\hat{w}_i = 1/\hat{p}_i,\ i \in S = S_1^* \cup S_2$
  - $\hat{t}_{y,joint} = \hat{N}\hat{\bar{y}} = \frac{n_1}{\hat{p}_1}\hat{\bar{y}} = \frac{n_1}{\sum_{i\in s_1\cap s_2} w_i / \sum_{i\in s_2} w_i} \frac{\sum_{i\in s_1\cup s_2} \hat{w} y_i}{\sum_{i\in s_1\cup s_2} \hat{w}_i}$
- Disjoint Weighting:
  - The samples are representative of the population, seperately
  - $\hat{w}_i = \begin{cases} 1/\hat{q}_i & i \in S_1^* \\ 1/\alpha_i & i \in S_2 \end{cases}$
  - $\hat{t}_{y,disjoint} = \hat{N}\hat{\bar{y}} = \frac{n_1}{\hat{p}_1}(\theta \frac{\sum_{i\in S_1^*} \hat{w}_i y_i}{\sum_{i\in S_1^*} \hat{w}_i} + (1-\theta)\frac{\sum_{iins_2} \hat{w}_i y_i}{\sum_{i\in S_2} \hat{w}_i})$
- From APSA method: $\hat{t}_{y,joint\_adp}$ and $\hat{t}_{y,disjoint\_adp}$

Jackknife Variance Estimation

- Segmentation:
  - $S_1 = S_1^{(1)} \cup S_1^{(2)} \cup \ldots \cup S_1^{(G)}$ *and* $S_2 = S_2^{(1)} \cup S_2^{(2)} \cup \ldots \cup S_2^{(G)}$
- For each replicate:
  - Leave "one" out from both samples
  - Calibration on the remaining samples
  - Re-fit PSA and APSA methods
- Jackknife variance estimator:

$$\widehat{Var}\left(\hat{\theta}\right) = \frac{G-1}{G}\sum_{g=1}^{G}(\hat{\theta}^{(g)} - \bar{\hat{\theta}})^2, \text{where } \bar{\hat{\theta}} = G^{-1}\sum_{g=1}^{G}\hat{\theta}^{(g)}.$$

Simulation Study

- Population: 2017 self-reported catch data from NMFS, 15771 trips
- Propensity score model: inclues all possible variables except response variable

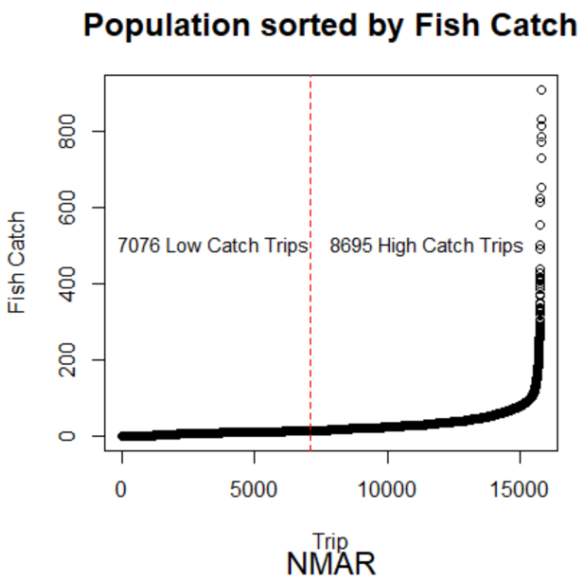
Variable Name	Type	Model Inclusion	Description
Trip ID	Cat.	No	Identification number of each trip
kept	Cont.	Yes	Fish catch by species
ReleasedAlive	Cont.	Yes	Alive fish released by species
ReleasedDead	Cont.	Yes	Dead fish released by species
TID	Cat.	No	Vessel Number
CaptainName	Cat.	No	Captain's name of the boat
Latitude	Cont.	No	Latitude when self-reported
Longitude	Cont.	No	Longitude when self-reported
NbPassengers	Cont.	Yes	Number of passengers on the boat
NbAnglers	Cont.	Yes	Number of anglers on the boat
NbCrew	Cont.	Yes	Number of Crew on the boat
Region	Cat.	Yes	Region of the boat: C I O
DepthPrimary	Cont.	Yes	Depth of the sea when fishing
Hours	Cont.	Yes	Fishing duration
State	Cat.	Yes	Belonging state of the boat: AL, FL, LA, MS, TX
County	Cat.	No	Belonging county of the boat
Name	Cat.	No	Name of the boat

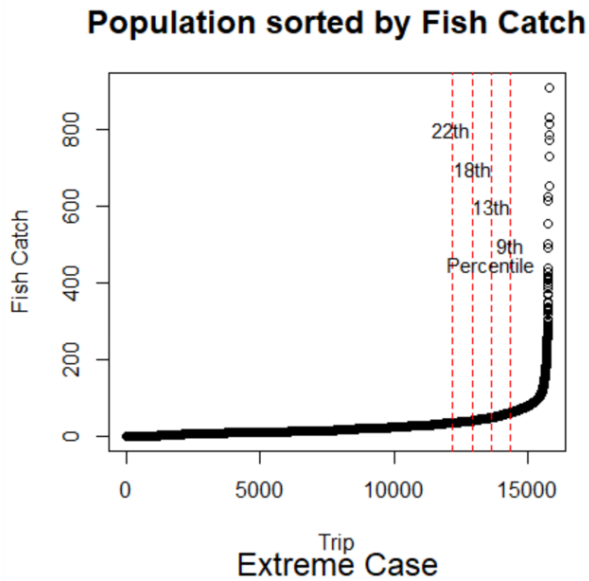
Simulation Settings

- Goal: evaluate $\hat{t}_{y,joint}$, $\hat{t}_{y,joint}$ from PSA, $\hat{t}_{y,joint\_adp}$, $\hat{t}_{y,disjoint\_adp}$ from APSA as alternatives to $\hat{t}_{MR}$
- 64 Settings based on 3 factors: 4 probability sample sizes $\times$ 4 non-probability sample sizes $\times$ 4 self-reporting mechanisms
- 5,000 replicates for each setting
- Probability sample: Dockside Intercept Sample
  - Simple Random Sample Design
  - Sample sizes: 200, 400, 600, 800
- Non-probability sample: Self-reported sample
  - Sample sizes: 3154, 4731, 6308, 7885
  - Corresponding reporing rate: 0.1, 0.2, 0.3, 0.4

Different Self-reporting Mechanisms

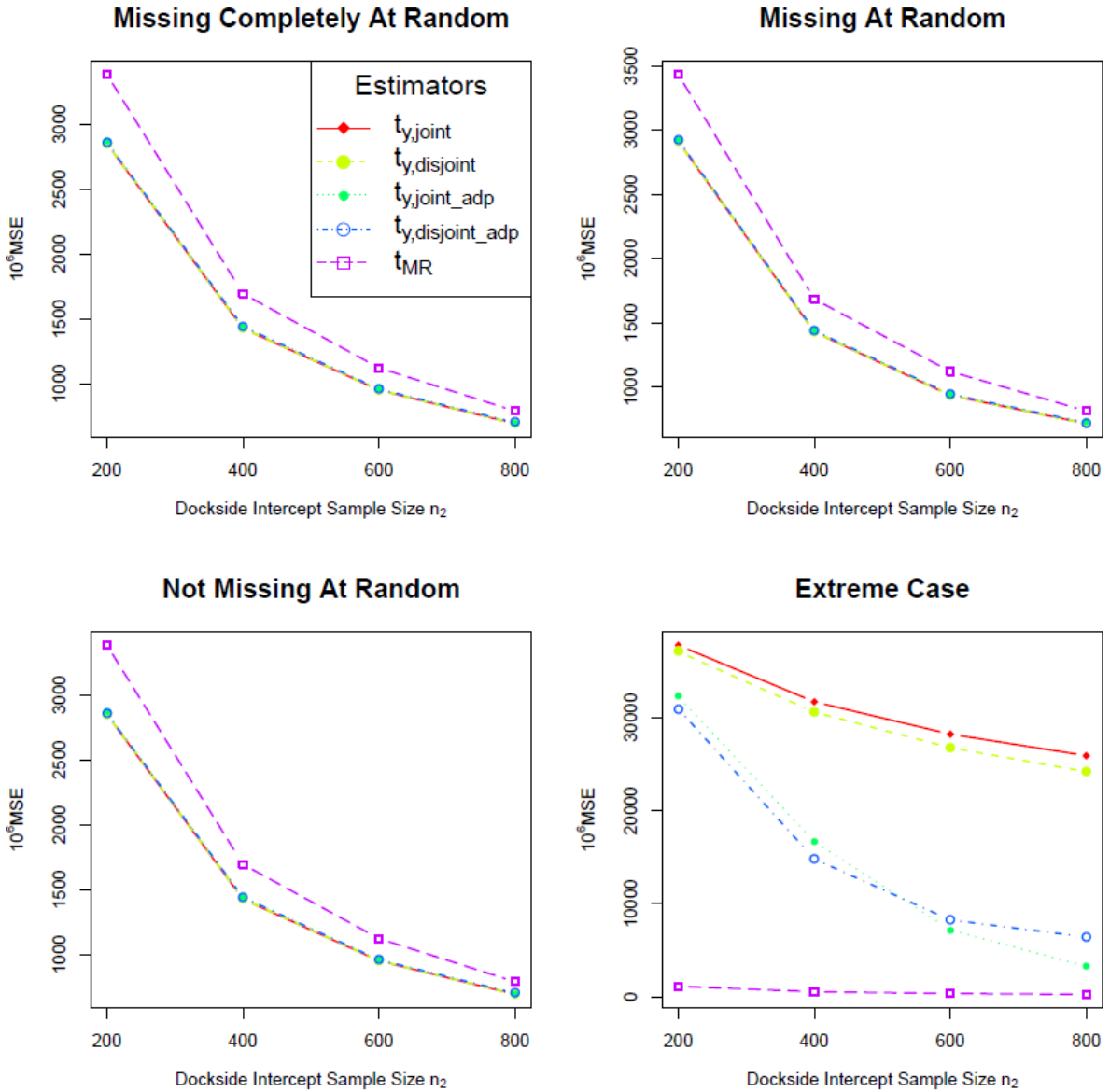
- a) Missing Completely at Random (MCAR)
  - simple random sample
- b) Missing at Random (MAR)
  - $log(\frac{\beta_i}{1-\beta_i}) = 0.5 \times NbPassengers + 0.5 \times NbCrew + 0.5 \times Hours + 1$
- c) Not Missing at Random (NMAR)
- d) Extreme Case





Simulation Result: Different Scenarios

MSE of Different Estimators by Scenarios: Self-reporting rate of 0.2



Simulation Result: Extreme Case

Simulation Results in Extreme Case

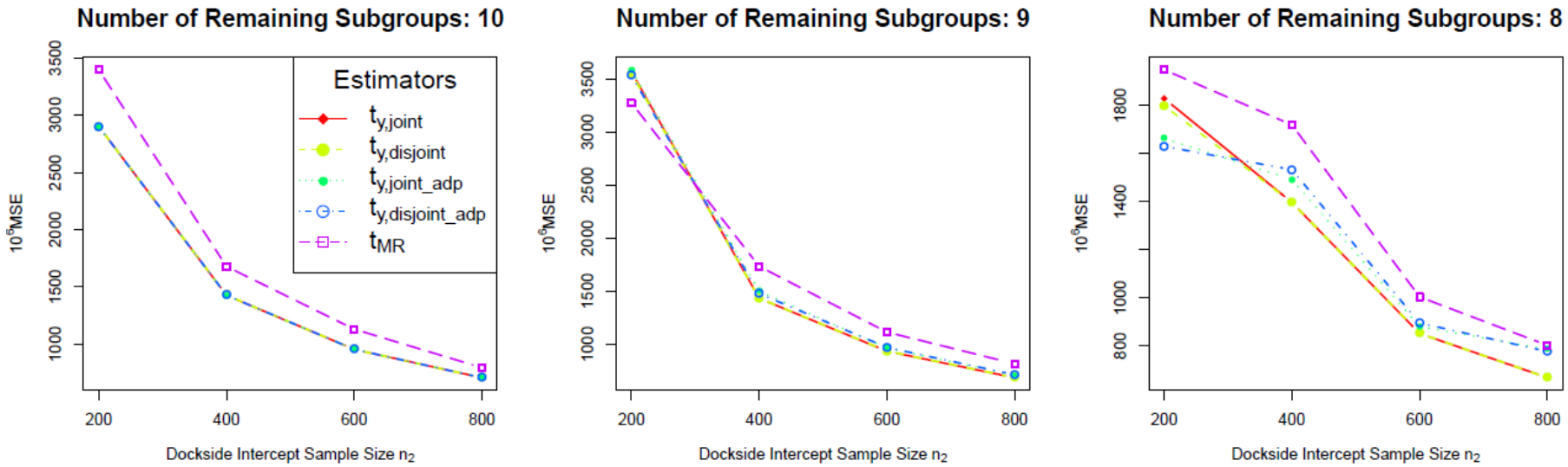
		PSA		APSA		Ratio Estimator
n_2		$\hat{t}_{y,joint}$	$\hat{t}_{y,disjoint}$	$\hat{t}_{y,joint\_adp}$	$\hat{t}_{y,disjoint\_adp}$	$\hat{t}_{MR}$
200	Bias ($\times 10^3$)	182.62	181.62	167.62	163.62	3.62
	Var ($\times 10^6$)	4273	4240	4232	4272	1100
	MSE ($\times 10^6$)	37803	37105	32290	30964	1110
400	Bias ($\times 10^3$)	172.62	169.62	113.62	83.62	1.62
	Var ($\times 10^6$)	1996	1971	3730	7835	523
	MSE ($\times 10^6$)	31723	30615	16684	14810	525
600	Bias ($\times 10^3$)	164.62	159.62	63.62	2.62	0.62
	Var ($\times 10^6$)	1215	1195	3133	8245	336
	MSE ($\times 10^6$)	28240	26794	7172	8253	337
800	Bias ($\times 10^3$)	158.62	152.62	36.62	-39.38	0.62
	Var ($\times 10^6$)	851	834	1925	4888	240
	MSE ($\times 10^6$)	25927	24206	3252	6413	241

Simulation Result: Coverage Rate from Jackknife Variance Estimation and Number of Remaining Subgroups from APSA

		Coverage Rate					#Subgroups
		PSA		APSA		Ratio Estimator	
	n_2	$\hat{t}_{y,joint}$	$\hat{t}_{y,disjoint}$	$\hat{t}_{y,joint\_adp}$	$\hat{t}_{y,disjoint\_adp}$	$\hat{t}_{MR}$	
MCAR	200	0.94	0.94	0.95	0.95	0.93	9.75
	400	0.94	0.94	0.95	0.95	0.93	9.74
	600	0.93	0.93	0.95	0.95	0.93	9.73
	800	0.94	0.94	0.96	0.96	0.94	9.74
MAR	200	0.94	0.94	0.95	0.95	0.93	9.75
	400	0.94	0.94	0.95	0.95	0.94	9.74
	600	0.94	0.94	0.95	0.96	0.93	9.75
	800	0.93	0.93	0.95	0.95	0.93	9.74
NMAR	200	0.94	0.94	0.95	0.95	0.93	9.75
	400	0.94	0.94	0.95	0.95	0.93	9.74
	600	0.93	0.93	0.95	0.95	0.93	9.73
	800	0.94	0.94	0.96	0.96	0.94	9.74
Extreme	200	0.07	0.07	0.48	0.53	0.94	6.20
	400	0.00	0.00	0.73	0.85	0.93	3.71
	600	0.00	0.00	0.88	0.95	0.93	2.78
	800	0.00	0.00	0.94	0.91	0.94	2.37

Simulation Result: Summarized by Number of Remaining Subgroups

MSE of Different Estimators by Number of Remaining Subgroups



Simulation Result: Coverage Rate of Different Number of Remaining Subgroups

		Coverage Rate				
		PSA		APSA		Ratio Estimator
#Subgroups	n_2	$\hat{t}_{y,joint}$	$\hat{t}_{y,\ disjoint}$	$\hat{t}_{y,joint\_adp}$	$\hat{t}_{y,disjoint\_adp}$	$\hat{t}_{MR}$
10	200	0.94	0.94	0.95	0.95	0.93
	400	0.94	0.94	0.95	0.95	0.93
	600	0.93	0.93	0.95	0.95	0.93
	800	0.94	0.94	0.95	0.95	0.94
9	200	0.92	0.92	0.94	0.94	0.93
	400	0.94	0.94	0.96	0.96	0.93
	600	0.93	0.93	0.95	0.96	0.93
	800	0.94	0.94	0.97	0.97	0.93
8	200	0.11	0.11	0.45	0.48	0.93
	400	0.45	0.45	0.77	0.77	0.84
	600	1.00	1.00	1.00	1.00	1.00
	800	1.00	1.00	1.00	1.00	1.00

Conclusions

- Both $\hat{t}_{y,joint}$, $\hat{t}_{y,joint}$ from PSA, $\hat{t}_{y,joint\_adp}$, $\hat{t}_{y,disjoint\_adp}$ from APSA have potential of being a useful alternative to $\hat{t}_{MR}$
- APSA can monitor the sample integration process and detect the non-representative part of the non-probability sample
- Compared to PSA, APSA can further reduce the selection bias by filtering out the non-representativeness part of the non-probability sample
- The performance of APSA will be improved by a larger probability sample
- Certains limits in adjusting selection bias in PSA and APSA
- Recommand to use PSA or APSA when the number of remaining subgroups greater than 8

Questions

- Is our approach unique?
  - Having variable of interest from both samples?
  - Conduct APSA on some covariate which are highly correlated with the variable of interest?
- How to conduct Jackknife Variance Estimation when the non-probability sample contains no design parameters
- Machine learing techniques other than propensity score?

Acknowledgement

Dr. Lynne Stokes
Professor and Chair
Department of Statistical Science
Southern Methodist University

References

- Elliott, M. R., and Valliant, R. (2017). Inference for nonprobability samples. Statistical Science, 32(2), 249-264.
- Baker, R., Brick, J. M., Bates, N. A., Battaglia, M., Couper, M. P., Dever, J. A., ... and Tourangeau, R. (2013). Summary report of the AAPOR task force on non-probability sampling. Journal of Survey Statistics and Methodology, 1(2), 90-143.
- Robbins, M. W. (2017, August). Blending of Probability and Convenience Samples as Applied to a Survey of Military Caregivers. In Joint Statistical Meetings.
- Lee, S., and Valliant, R. (2009). Estimation for volunteer panel web surveys using propensity score adjustment and calibration adjustment. Sociological Methods and Research, 37(3), 319-343.
- Lee, S. (2006). Propensity score adjustment as a weighting scheme for volunteer panel web surveys. Journal of official statistics, 22(2), 329.