Exchangeability Assumption in Propensity-Score Based Adjustment Methods for Population Mean Estimation Using Non-Probability Samples

Yan Li

Joint Program in Survey Methodology &

Department of Epidemiology and Biostatistics,

University of Maryland at College Park

3/1/22

Morris Hansen Lecture

This work has been generously supported by NIH-R03 CA252782

“Improving Population Representativeness of the Inference from Non-Probability Sample Analysis,” National Institute of Health

This work is an extension of two papers

- L. Wang, B.I. Graubard, H.A. Katki, Y. Li (2021). Efficient and Robust Propensity-Score-Based Methods for Population Inference using Epidemiologic Cohorts. *International Statistical Review*.
- L. Wang, B.I. Graubard, H.A. Katki, Y. Li (2020). Improving external validity of epidemiologic cohort analyses: a kernel weighting approach, *Journal of Royal Statistical Society A*, 183, 1293-311.

Population Inference using Nonprobability Samples

- Nonprobability samples subject to Selection Bias
- Common Approaches for Improving the population representation
  - Model-based Methods
    - Regression (Wang et al. 2015)
  - Propensity Score (PB)-based adjustment
    - PS *Weighting* (Wang, et al. 2021; Chen, et al. 2020; Elliott and Valliant, 2017; Kim, et al. 2018, Rafei et al. 2020; etc.)
    - PS *Matching* (Valliant and Lee 2010; River, 2007; Wang, et al. 2020; Wang, et al. 2021; Yang et al. 2021; etc)
  - Doubly Robust
- Review Paper: Beaumont (2021); Rao (2021); Valliant (2020); Yang and Kim (2020)

Assumptions

- PS-based methods
  - Propensity model
  - **Conditional Exchangeability**
  - Positivity
  - Representative probability sample
  - etc…
- Model-based method
  - Outcome model
  - Transportability
  - etc…

Notation

- Y: Outcome variable of interest
- ***X***: a vector of observed covariates

- U: the set of the finite population units of size N
- C: the set of the nonprobability sample units and $C \subset U$

- **Challenge**: We observe C, which is NOT representing *U*

$$E_C(y) \neq E_U(y)$$

Estimating $E(y|U)$

- Assume Conditional Exchangeability

$$E_C\{y|b(\boldsymbol{x})\} = E_U\{y|b(\boldsymbol{x})\}, \qquad (*)$$

where

$b(\boldsymbol{x})$: a function of covariates $\boldsymbol{x}$, called balancing score

- Choices of the balancing score
  - **Basic criteria:** Distinguish *C* units by participation rates
  - A natural choice: $b(\boldsymbol{x}) = P(i \in C|\, \boldsymbol{x}, U)$
  - Other choices: **Finer than, if not equal to, $P(i \in C|\, x, U)$**
    - *Finest* balancing score: $b(\boldsymbol{x}) = \boldsymbol{x}$
    - *Coarsest:* $b(\boldsymbol{x}) = P(i \in C|\, \boldsymbol{x}, U)$ or its monotone function (Rosenbaum and Rubin, 1983)

Estimation of $\boldsymbol{p}(\boldsymbol{i} \in \boldsymbol{C} | \boldsymbol{x}, \boldsymbol{U})$

- *S*: the set of a reference probability sample units with $\{\boldsymbol{x}_i : i \in S\}$
- Various parametric or nonparametric models, e.g.,

$$\log\left\{\frac{p(\boldsymbol{x}_i)}{1-p(\boldsymbol{x}_i)}\right\} = \boldsymbol{B}^T g(\boldsymbol{x}_i), \qquad \text{for } i \in C \cup S, \qquad (1)$$

  - $p(\boldsymbol{x}_i)$: likelihood of being units in C vs. U, and

$$P(i \in C | x, U) = \exp\left(\boldsymbol{B}^T g(\boldsymbol{x}_i)\right)$$

  - $g(\boldsymbol{x}_i)$ is a known function of observed covariates
  - $\boldsymbol{B}$ the unknown regression coefficients

- $\widehat{\boldsymbol{B}}_w$: Estimated by fitting (1) to combined C and weighted S

- Define $b\left(\boldsymbol{x}; \widehat{\boldsymbol{B}}_w\right) = \widehat{\boldsymbol{B}}_{\boldsymbol{w}}^{\boldsymbol{T}} g(\boldsymbol{x}_i) = \mathbf{log}\, P(i \in C | x_i, U)$. Therefore,

$$E_C\left\{y \middle| b\left(\boldsymbol{x}; \widehat{\boldsymbol{B}}_w\right)\right\} = E_U\left\{y \middle| b\left(\boldsymbol{x}; \widehat{\boldsymbol{B}}_w\right)\right\}$$

PS-based Adjustment Estimators

- PS-Weighting: Weight units in C by inverse of $\exp\left(b\left(\boldsymbol{x}, \widehat{\boldsymbol{B}}_w\right)\right)$
- PS-Matching: Match units in C and S based on $b\left(\boldsymbol{x}; \widehat{\boldsymbol{B}}_w\right)$

Properties

- Approximately unbiased (Wang et al. 2020; 2021)
- **Challenge**: Variance Inflation – sample weights in C vs. S (Scott and Wild, 1986)

QUESTION: Estimate $\boldsymbol{B}$ ignoring survey weights in (1), $\widehat{\boldsymbol{B}}_0$,

$$\text{Define } b\left(\boldsymbol{x};\widehat{\boldsymbol{B}}_0\right)=\widehat{\boldsymbol{B}}_{\mathbf{0}}^{\boldsymbol{T}}g(\boldsymbol{x}_i)$$

$$\text{Is } E_C\left\{y\middle|b\left(\boldsymbol{x};\widehat{\boldsymbol{B}}_0\right)\right\}=E_U\left\{y\middle|b\left(\boldsymbol{x};\widehat{\boldsymbol{B}}_0\right)\right\}?$$

Let us think:

- $b\left(\boldsymbol{x};\widehat{\boldsymbol{B}}_0\right)$ produces sample balance in $\boldsymbol{x}$ between C and S

  $$x\perp(C,S)|b(\boldsymbol{x};\widehat{\boldsymbol{B}}_0)$$

  and therefore

  $$E_{\boldsymbol{C}}\left\{y\middle|b\left(\boldsymbol{x};\widehat{\boldsymbol{B}}_0\right)\right\}=E_{\boldsymbol{S}}\left\{y\middle|b\left(\boldsymbol{x};\widehat{\boldsymbol{B}}_0\right)\right\}$$

- IS $E_{\boldsymbol{C}}\left\{y\middle|b\left(\boldsymbol{x};\widehat{\boldsymbol{B}}_0\right)\right\}=E_{\boldsymbol{U}}\left\{y\middle|b\left(\boldsymbol{x};\widehat{\boldsymbol{B}}_0\right)\right\}$? Equivalently, Is $b\left(\boldsymbol{x};\widehat{\boldsymbol{B}}_0\right)$ a finer or monotone function of $b\left(\boldsymbol{x};\widehat{\boldsymbol{B}}_w\right)$? E.g. $\widehat{\boldsymbol{B}}_0=const.\widehat{\boldsymbol{B}}_w$.

GOOD LUCK!

An Adaptive Exchangeability Assumption

- 1st step – Fit the combined sample $C \cup S$ to

$$\log\left\{\frac{p(i \in C)}{p(i \in S)}\right\} = \alpha + \boldsymbol{B}^T g(\boldsymbol{x}_i), \qquad \text{for } i \in C \cup S$$

$$\rightarrow b\left(\boldsymbol{x}; \widehat{\boldsymbol{B}}_0\right) = \widehat{\boldsymbol{B}}_{\mathbf{0}}^T g(\boldsymbol{x}_i)$$

- *2nd step* – Fit the combined sample $S \cup S_w$ to

$$\log\left\{\frac{p(i \in S)}{p(i \in S_w)}\right\} = \gamma_0 + \boldsymbol{\gamma}^T g(\boldsymbol{x}_i), \qquad \text{for } i \in S \cup S_w$$

$$\rightarrow b(\boldsymbol{x}; \widehat{\boldsymbol{\gamma}}_w) = \widehat{\boldsymbol{\gamma}}_w^T g(\boldsymbol{x}_i)$$

- 3rd step – Construct the new balancing score by adding them up

$$b'(\boldsymbol{x}) = \log\left\{\frac{p(i \in C)}{p(i \in S_w)}\right\} = \left(\widehat{\boldsymbol{\gamma}}_w^T + \widehat{\boldsymbol{B}}_0^T\right) g(\boldsymbol{x}_i), \qquad \text{for } i \in C \cup S$$

PS matching based on $\boldsymbol{b}'(\boldsymbol{x})$

e.g., Kernel Weighting (KW) method by Wang et al. JRSS A 2020

$$w_j^{kw} = \sum_{i \in S} w_i \left(\frac{K\left(\frac{d_{ij}}{h}\right)}{\sum_{j \in C} K\left(\frac{d_{ij}}{h}\right)} \right) \text{ for } j \in C$$

- w_i is the sample weight of survey unit i
- $K(\cdot)$ is an arbitrary kernel function such as standard normal
- h is the bandwidth associated with $K(\cdot)$
- $d_{ij} = b'(\boldsymbol{x_i}) - b_0{}'(\boldsymbol{x_j})$

$$\bar{y}^{kw} = \frac{\sum_{j \in C} w_j^{kw} y_j}{\sum_{j \in C} w_j^{kw}}$$

SIMULATION STUDIES

Finite population generation U

- N=120,000

- Three covariates x_1, x_2, x_3 ~ $N(0,1)$ with pairwise correlations $\rho_{x_1x_3} = \rho_{x_2x_3} = 0$ and $\rho_{x_1x_2} = 0.2$

- Binary outcome Y with varying α_0 with prevalence of 29%, 15% or 7%

$$P(Y=1) = \frac{\exp\left(\alpha_0 + x_1\alpha_{x_1} + x_2\alpha_{x_2} + x_1x_2\alpha_{x_1x_2}\right)}{1+\exp\left(\alpha_0 + x_1\alpha_{x_1} + x_2\alpha_{x_2} + x_1x_2\alpha_{x_1x_2}\right)}$$

Outcome predictors: $\boldsymbol{x_1}$ and $\boldsymbol{x_2}$

Probability Sample (*S*) & Non-probability Sample (*C*) Selection

- n_S =500 and n_C =1500
- Probability proportional to size sampling with measure of size

$$MOS = \exp\left(a \times \boldsymbol{\beta}^{\boldsymbol{T}} \boldsymbol{x}\right)$$

- Probability Samples with $\boldsymbol{x} = (\boldsymbol{x_1}, \boldsymbol{x_3})$ in MOS
  - Vary CV(weights) by setting a = 0.1, 0.5, 1 or 2
- Nonprobability samples – Unknown underlying selection process
  - Quota sample on joint distributions of both $\boldsymbol{x_1}$ and $\boldsymbol{x_2}$
  - Quota sample on distribution of $\boldsymbol{x_1}$ or $\boldsymbol{x_2}$
  - Volunteer sample with unbalanced distributions in both $\boldsymbol{x_1}$ and $\boldsymbol{x_2}$

PS matching estimators of population mean

- KW with $b\left(\boldsymbol{x}; \widehat{\boldsymbol{B}}_w\right)$ – Approx. unbiased but inflated variance
- KW with $b\left(\boldsymbol{x}; \widehat{\boldsymbol{B}}_0\right)$ – Can be biased but more efficient
- KW with $b'(\boldsymbol{x})$ – Approx. unbiased with reduced variance

Evaluation Criteria

- RelBias (%) = (mean of 300 simulated means - population mean) divided by population mean $\times 100\%$
- EmpVar ($\times 10^4$) = variance of 300 simulated means
- MSE ($\times 10^4$) = square of bias + empirical variance

Results

1. Reference survey: (close to) self-weighted

$$b\left(\boldsymbol{x};\widehat{\boldsymbol{B}}_0\right)\approx \boldsymbol{b}'(\boldsymbol{x})\approx b\left(\boldsymbol{x};\widehat{\boldsymbol{B}}_w\right) \text{ due to } b(\boldsymbol{x};\widehat{\boldsymbol{\gamma}}_w)\approx \mathbf{0}$$

2. Reference survey: variable weights

a. Quota sample on joint distribution of x_1 and x_2

$$b\left(\boldsymbol{x};\widehat{\boldsymbol{B}}_0\right)\approx \boldsymbol{b}'(\boldsymbol{x}) \textbf{ more efficient than } b\left(\boldsymbol{x};\widehat{\boldsymbol{B}}_w\right)$$

Quota sample with balanced distribution in outcome predictors

	Probability samples PPS($\boldsymbol{MOS}$)			
	$a = 0.1$	$a = 0.5$	$a = 1$	$a = 2$
CV.wts	0.07	0.38	0.86	2.29
	RelBias(%)			
$b(\boldsymbol{x}; \widehat{\boldsymbol{B}}_w)$	0.34	0.00	0.36	1.45
$b'(\boldsymbol{x})$	0.34	0.00	0.00	-0.36
	EmpVar			
$b(\boldsymbol{x}; \widehat{\boldsymbol{B}}_w)$	1.94	2.18	3.08	6.95
$b'(\boldsymbol{x})$	1.94	2.11	2.69	3.53
	MSE			
$b(\boldsymbol{x}; \widehat{\boldsymbol{B}}_w)$	1.94	2.18	3.08	7.07
$b'(\boldsymbol{x})$	1.95	2.11	2.69	3.53

b. Quota sample on a subset of predictors, x_1, not in x_2 and x_3

	Probability samples with PPS$(\boldsymbol{MOS})$			
	$a = 0.1$	$a = 0.5$	$a = 1$	$a = 2$
CV.wts	0.07	0.38	0.86	2.29
	RelBias(%)			
$b(\boldsymbol{x}; \widehat{\boldsymbol{B}}_w)$	0.34	0.36	0.36	1.45
$b(\boldsymbol{x}; \widehat{\boldsymbol{B}}_0)$	2.05	10.18	13.09	5.82
$b'(\boldsymbol{x})$	0.34	0.36	0.36	0.36
	EmpVar			
$b(\boldsymbol{x}; \widehat{\boldsymbol{B}}_w)$	2.28	2.31	2.79	8.87
$b(\boldsymbol{x}; \widehat{\boldsymbol{B}}_0)$	2.35	2.31	2.11	4.65
$b'(\boldsymbol{x})$	2.27	2.13	2.35	4.05
	MSE			
$b(\boldsymbol{x}; \widehat{\boldsymbol{B}}_w)$	2.29	2.32	2.80	9.01
$b(\boldsymbol{x}; \widehat{\boldsymbol{B}}_0)$	2.70	9.99	15.01	7.26
$b'(\boldsymbol{x})$	2.27	2.14	2.35	4.06

Real Data Analysis

1. COVID with BRFSS as reference (Kalish et al. 2021)

2. Unweighted NHANES with NHIS as reference (Wang et al. 2021)

Data Example I – NIH SARS-CoV-2 seroprevalence study

AIM: Proportion of U.S. adults with COVID-19 antibodies from April 01 to August 04, 2020

NIH SARS-CoV-2 seroprevalence study (Kalish et al., 2021)

- More than 460,000 volunteers responding within weeks of the study announcement
- Select subset of volunteers based on *age, race, sex, ethnicity and region*
- A sample of 8058 subjects answered a questionnaire on medical, geographic, demographic, and socioeconomic information and provided blood samples
- Quota Sampling - Rapid data collection but suffer from **Selection Bias**

Behavioral Risk Factor Surveillance System (BRFSS) survey (CV(wt) = 1.92)

- A national representative probability survey
- Adjust for potential selection bias by 11 variables related to seropositivity but were not used in the quota sampling
- A total of 367,165 participants, responded to the same clinical questionnaire, were included in the analysis

		Covid Survey	Weighted BRFSS
Age Group			
	18-45	41.6	42.9
	45-70	42.6	41.8
	70-95	15.8	15.2
Sex			
	Male	47.4	47.8
	Female	52.6	52.2
Race			
	White only	77.5	74.8
	Black only	9.4	12.6
	Others	13.1	12.5
Ethnicity			
	Hispanic	15.9	14.1
	Not Hispanic	84.1	85.9
Region			
	Northeast	16.7	17.1
	Midwest	15.8	17.6
	Mid-Atlantic	20.8	17.3
	South/Central	14.2	15.7
	Mountain/Southwest	15.5	15.3
	West/Pacific	17.0	16.9

		Covid Survey	Weighted BRFSS
Urban/Rural			
	Urban	94.7	93.2
	Rural	5.3	6.8
Children present			
	Yes	32.5	34.7
	No	67.5	65.3
Educ3			
	<=HS	2.6	39.4
	College	13.8	31.5
	>=College	83.6	29.1
Homeowner			
	Own	75.2	68.8
	Rent	20.2	25.6
	Others	4.7	5.6
Employment			
	Employed	71.2	57.4
	NLF	23.8	32.2
	Unemployed	5.0	10.4

		Covid Survey	Weighted BRFSS
Flu Vaccinated			
	Yes	73.8	51.3
	No	26.2	48.7
Cardiovascular			
	Yes	4.1	9.5
	No	95.9	90.5
Pulmonary			
	Yes	18.8	18.7
	No	81.2	81.3
Immune			
	Yes	23.4	31.1
	No	76.6	68.9
Diabetes			
	Yes	5.5	11.9
	No	94.5	88.1
Health Insurance			
	Yes	97.4	89.0
	No	2.6	11.0

Undiagnosed seropositivity rate among US adults 04/01/2020-08/04/2020

KW Matching	est (%)	se* ($\times 10^{-2}$)
$b(\boldsymbol{x}; \widehat{\boldsymbol{B}}_w)$	6.79	**2.50**
$b(\boldsymbol{x}; \widehat{\boldsymbol{B}}_0)$	4.32	0.66
$b'(\boldsymbol{x})$	4.31	0.67
Post-stratification		
$b(\boldsymbol{x}; \widehat{\boldsymbol{B}}_w)$	4.56	0.83
$b(\boldsymbol{x}; \widehat{\boldsymbol{B}}_0)$	4.39	0.61
$b'(\boldsymbol{x})$	4.33	0.61

*: no account for the variability due to estimating B or $\boldsymbol{\gamma}$

Data Example II -- NHANES III & NHIS 1994

~ Estimate prospective 15-year all-cause mortality for people aged 18 to 75 in the US from 1990

- The Third National Health and Nutrition Examination Survey (NHANES)
  $n_c = 17{,}111, \quad \hat{N} = 173{,}481{,}294$
- Reference Survey: 1994 National Health Interview Survey (NHIS)
  $n_s = 18{,}138, \quad \hat{N} = 178{,}226{,}524$ and CV(NHIS weights) = 0.57

Both Surveys oversample **old people (>= 60 yrs)**, minorities, low-income

Note: The two surveys share target population, data collection mode, well-designed questionnaires, and mortality information Linked to NDI.

NHIS-weighted 15-year all-cause mortality=13.04%

Estimate of 15-year Mortality Rate (%) using unweighted NHANES

	NHIS	NHANES	$b'(\boldsymbol{x})$	$b\left(\boldsymbol{x}; \widehat{\boldsymbol{B}}_0\right)$	$b\left(\boldsymbol{x}; \widehat{\boldsymbol{B}}_w\right)$
Full Sample	**13.0**	17.9	**13.5%**	16.0	13.4%
[18,30]	2.1	2.5	**2.3%**	2.3	2.3%
(30,50]	6.0	7.5	**5.0%**	5.6	5.0%
(50,75]	34.6	41.7	**35.5%**	37.8	35.5%

	Propensity of Unweighted NHIS vs. Weighted NHIS		Logistic Regression of Outcome	
	Estimate	pvalue	Estimate	pvalue
age_c2	0.202	0.000	1.057	0.000
age_c3	0.230	0.000	3.071	0.000
Sex	0.175	0.000	-0.573	0.000
Educ6	0.051	0.000	-0.065	0.002
race2	0.014	0.860	-0.032	0.597
race3	-0.094	0.417	-0.554	0.000
race4	-0.171	0.143	-0.855	0.001
Poverty	-0.123	0.023	-0.232	0.002
poverty3	-0.144	0.051	-0.064	0.424
Health	-0.020	0.137	0.386	0.000
region2	0.027	0.911	-0.040	0.624
region3	-0.059	0.798	0.018	0.801
region4	0.006	0.983	0.080	0.343
Marstat	0.450	0.000	0.294	0.000
marstat3	0.227	0.000	0.028	0.752
smk_stat1	0.003	0.935	0.648	0.000
smk_stat2	0.031	0.306	0.401	0.000
fam_inc	-0.084	0.000	-0.164	0.000
snuff_chew	0.003	0.946	0.104	0.212

Conclusion and Discussion

- Conditional Exchangeability (*) - balancing scores **Finer** than, if not equal to, the participating rate
  - Weighted propensity scores $b\left(\boldsymbol{x}; \widehat{\boldsymbol{B}}_w\right)$
  - Unweighted propensity scores $b\left(\boldsymbol{x}; \widehat{\boldsymbol{B}}_0\right)$
- Adaptive exchangeability
  - Identify $b\left(\boldsymbol{x}; \widehat{\boldsymbol{B}}_0\right)$
  - Identify *bias correction factor* $b(\boldsymbol{x}; \widehat{\boldsymbol{\gamma}}_w)$ *by comparing S vs* S_w,
  - Construct $b'(\boldsymbol{x}) = b\left(\boldsymbol{x}; \widehat{\boldsymbol{B}}_0\right) + \boldsymbol{b}(\boldsymbol{x}; \widehat{\boldsymbol{\gamma}}_{\boldsymbol{w}})$**,** which a monotone function of $P(i \in C|\boldsymbol{x}, U)$.

Future Area

- Other methods to satisfy adaptive exchangeability? Poststratification?
- Variables to be collected in both C and S?
- Propensity Modeling and Estimation
  - Depends on the predictivity of propensity score model?
  - Machine learning Methods?

- High quality reference survey required by $\boldsymbol{b}(\boldsymbol{x}; \widehat{\boldsymbol{\gamma}}_{\boldsymbol{w}})$

Less variable and informative weights!

High-Quality Probability Samples are still in great demand, especially for population-level descriptive estimates

THANK YOU!

REFERENCES

- Beaumont, J.-F. (2020). Are probability surveys bound to disappear for the production of official statistics? *Survey Methodology*, 46, 1-28.
- Chen, Y., Li P., and Wu C. (2020). Doubly robust inference with nonprobability survey samples. Journal of the American Statistical Association 115, 2011-2021.
- Elliott M.R. and Valliant R. (2017). Inference for nonprobability samples. *Stat. Sci.*, 32, 249-64.
- Kalish, H., Klumpp-Thomas, C., Hunsberger, etc (2021). Mapping a Pandemic: SARS-CoV-2 Seropositivity in the United States. *medRxiv:* https://doi.org/10.1101/2021.01.27.21250570
- Kim, J. K., S. Park, Y. Chen, and C. Wu (2018). Combining non-probability and probability survey samples through mass imputation. arXiv preprint arXiv:1812.10694.
- Lee S. and Valliant R. (2009). Estimation for volunteer panel web surveys using propensity score adjustment and calibration adjustment. *Sociological Methods & Research*, 37, 319-43.
- Rafei, A., C. A. Flannagan, and M. R. Elliott (2020). Big data for finite population inference: Applying quasi-random approaches to naturalistic driving data using Bayesian additive regression trees. Journal of Survey Statistics and Methodology 8 (1), 148-180.
- Rao, J.N.K. (2021). On Making Valid Inferences by Integrating Data from Surveys and Other Sources. *Sankhya* B, 83:242-272.

- Rivers D. (2007). Sampling for web surveys. Paper presented at the Joint Statistical Meetings, Section on Survey Research Methods. Salt Lake City, Utah.
- Rosenbaum P.R. and Rubin D.B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika, 70***,** 41-55.
- Valliant (2020). Comparing alternatives for estimation from nonprobability samples. *Journal of Survey Statistics and Methodology*, 8, 231-263.
- Wang, L., Valliant, R., and Li, Y. (2021). Adjusted Logistic Propensity Weighting Methods for Population Inference using Nonprobability Volunteer-Based Epidemiologic Cohorts. *Stat. in Med.,* https://doi.org/10.1002/sim.9122.
- Wang, L., Graubard, B.I., Hormuzd, A.K. and Li, Y. (2021). Efficient and Robust Propensity-Score-Based Methods for Population Inference using Epidemiologic Cohorts. *International Statistical Review* (accepted).
- Wang, W., Rothschild, D.; Goel, S.; Gelman, A. (2015). "Forecasting elections with non-representative polls" (PDF). *International Journal of Forecasting*. **31** (3): 980–991. doi:10.1016/j.ijforecast.2014.06.001.
- Yang, S., and Kim, J. K. (2020). Statistical data integration in survey sampling: A review. *Japanese Journal of Statistics and Data Science*, 1-26.
- Yang, S., Kim, J.K. and Hwang, Y. (2021). Integration of data from probability surveys and big found data for finite population inference using mass imputation. *Survey Methodology*, 47, 29-58.