Variable inclusion strategies for adjusting the weights of surveys subject to selection bias

Katherine Irimata, PhD, National Center for Health Statistics
Yan Li, PhD, University of Maryland
Yulei He, PhD, National Center for Health Statistics
Jennifer Parker, PhD, National Center for Health Statistics

2022 FCSM Research and Policy Conference
October 26, 2022

Population estimation from web-based surveys

- Web-based surveys, nonprobability and probability-sampled, can be used for more timely and cost-effective data collections
- However, these surveys may be subject to lower coverage and response rates than large nationally representative surveys
- Selection bias has been a concern due to differences in the composition of web panels compared to the total population, which can impact population mean estimation
- To adjust for these differences, weighting methods have been applied to web surveys to align the covariate distribution to a high-quality benchmark

Propensity score-adjustment methods

- Propensity score (PS) methods were developed by Rosenbaum and Rubin (1983) to control for confounding in treatment estimation in observational studies
- In survey research, PS-based adjustment methods are used as a reweighting method to align the distribution of specified variables between a target (web) survey and a high-quality reference survey
- Estimated propensity scores are incorporated into the weights using various approaches such as PS weighting and PS matching

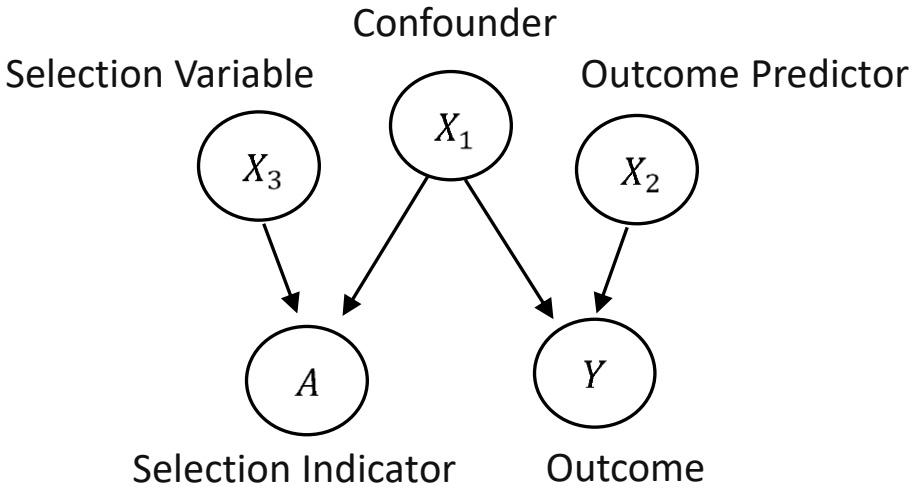
Covariate inclusion in propensity score models

- Some literature recommends including all variables collected in both the target and reference data sources in PS-adjustment
- Key question for constructing the pseudo-weights is which variables to include in the PS model to improve population mean estimation
- Study assesses the impact of selected covariates in PS-models on the bias and variance of the estimated population mean

Methods

Covariate types in PS adjustment

- Directed acyclic graph (DAG) used to examine how different variables in the causal pathways impact the performance of PS-adjustment



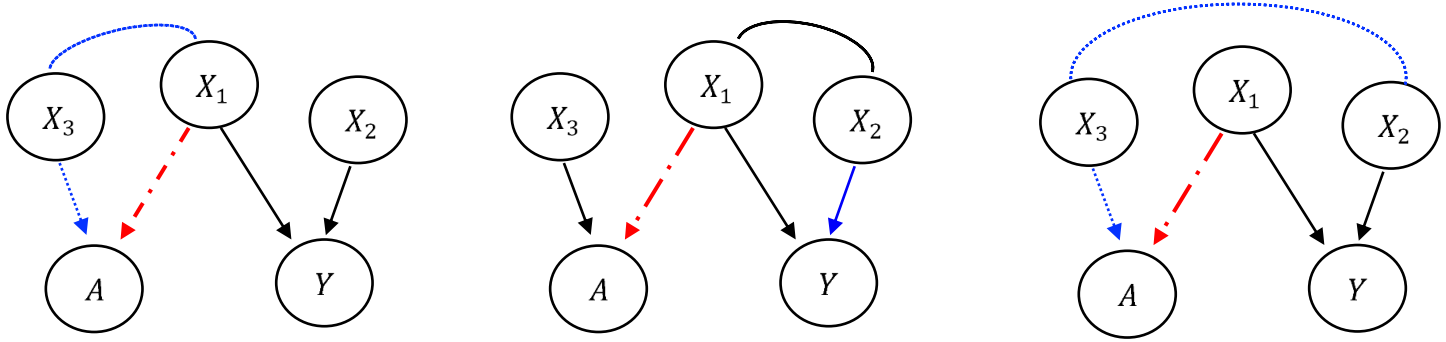
- In practice, any pairs or all confounders, outcome predictors, and selection variables may be correlated in the underlying population

Uncorrelated case: assessing bias and variance of population mean

- Numerically, it can be shown that confounders (X_1) induce bias in the estimate of the population mean (bias ≠ 0) if they are not balanced between the target sample and population
- In addition, it can be shown that the inclusion of selection variables (X_3) result in larger variance estimates since they are non-informative of the outcome Y
- **Propensity models should include confounders (X_1) alone or confounders (X_1) and outcome predictors (X_2) to produce unbiased and efficient mean estimates**
- **The inclusion of selection variables (X_3) in the propensity model does not add bias, but inflates the variance of the estimates**

Correlated case: assessing bias and variance of population mean

- Backdoor criteria (Pearl 2009) removes confounding via conditioning on a set of covariates that block the backdoor paths between A and Y



- **When (X_1 and X_3) or (X_1 and X_2) are correlated, X_1 should be included in the propensity model to produce an unbiased estimate of the mean**
- **When (X_2 and X_3) are correlated, (X_1 and X_2) or (X_1 and X_3) should be included in the propensity score model to produce unbiased estimates**

Simulation

Set up

Finite population (N=20,000)

- Three covariates (X_1, X_2, and X_3) simulated using trivariate normal distributions with specified pairwise correlations
- Binary outcome Y ~ Bernoulli distribution as a function of X_1 and X_2

Target Sample (N=1,000)

- Sample (A=1) selected from population using probability proportional to size (PPS) sampling with measure of size as a function of X_1 and X_3
- Inclusion probabilities (sample weights) are treated as unknown

Probability Sample (N=500)

- Sample (A=0) is selected using the same sampling design as the target sample selection with known selection probabilities

Conditions

Conducted over 500 iterations:

- **Simulation 1**: independent covariates in the finite population ($\rho_{x_1x_2} = \rho_{x_1x_3} = \rho_{x_2x_3} = 0$), weights adjusted using the PS matching method kernel weighting (KW, Wang et al. 2020)
- **Simulation 2**: varies covariate correlation in the finite population ($(\rho_{x_1x_2}, \rho_{x_1x_3}, \rho_{x_2x_3}) = (.6,0,0), (0,.6,0), (0,0,.6), (.6,.6,0), (.6,0,.6), (0,.6,.6),$ or $(.6,.6,.6)$), weights adjusted using KW
- **Simulation 3**: varies covariate correlation and includes interaction effects between covariates on the outcome and target sample inclusion ($\alpha_{12} = \beta_{13} = 0.5$), weights adjusted using KW and the PS weighting method adjusted logistic propensity (ALP)

Simulation 1 results

	Sample	w(x1)	w(x2)	w(x3)	w(x12)	w(x13)	w(x23)
Bias ($\times 10^2$)	4.61	0.26	4.50	4.83	0.26	0.41	4.77
Empirical Variance ($\times 10^4$)	2.20	2.68	2.62	2.96	2.92	3.43	3.32
Mean Squared Error ($\times 10^4$)	23.48	2.75	22.85	26.31	2.99	3.60	26.04

- Propensity models including the confounder (X_1) produce approximately unbiased estimates of the finite population mean of Y
- Propensity models containing the selection variable (X_3) result in inflated variance estimates
- Among the unbiased estimators, w(x1) yields the most efficient estimates

Simulation 2 results

	Sample	w(x1)	w(x2)	w(x3)	w(x12)	w(x13)	w(x23)
$(\rho_{x_1x_2}, \rho_{x_1x_3}, \rho_{x_2x_3}) = (.6, 0, 0)$							
Bias ($\times 10^2$)	7.35	0.37	2.98	7.60	0.37	0.59	3.25
Empirical Variance ($\times 10^4$)	2.15	2.59	2.64	2.77	2.66	2.88	2.84
Mean Squared Error ($\times 10^4$)	56.14	2.72	11.52	60.57	2.79	3.23	13.42
$(\rho_{x_1x_2}, \rho_{x_1x_3}, \rho_{x_2x_3}) = (0, .6, 0)$							
Bias ($\times 10^2$)	7.27	0.32	7.16	3.21	0.30	0.41	3.12
Empirical Variance ($\times 10^4$)	2.17	3.60	2.39	3.68	3.53	4.05	3.66
Mean Squared Error ($\times 10^4$)	54.98	3.70	53.67	13.97	3.62	4.22	13.39

- When (X_1 and X_3) or (X_1 and X_2) are correlated, PS-adjusted weights that balance the distributions of the confounders (X_1) produce approximately unbiased estimates

Simulation 2 results

	Sample	w(x1)	w(x2)	w(x3)	w(x12)	w(x13)	w(x23)
$(\rho_{x_1x_2}, \rho_{x_1x_3}, \rho_{x_2x_3}) = (0, 0, .6)$							
Bias ($\times 10^2$)	7.55	2.98	4.65	4.87	0.26	0.37	4.83
Empirical Variance ($\times 10^4$)	2.01	2.52	2.38	2.57	2.66	2.75	2.66
Mean Squared Error ($\times 10^4$)	59.00	11.38	24.00	26.30	2.73	2.89	26.03

- When correlation exists between X_2 and X_3 , inclusion of only the confounder (X_1) in the PS model induces bias
- For all correlation conditions, PS-adjusted weights that balance the distributions in the outcome predictors (X_2) or selection variables (X_3) along with the confounders (X_1) produce approximately unbiased estimates
- Empirical variance estimates and MSEs for models including X_1 and X_2 (w(x12)) tend to be smaller than models including X_1 and X_3 (w(x13))

Simulation 3 results

- ALP approach (PS weighting) yields unbiased estimates only under the true propensity model (model containing X_1, X_3, and interaction term X_1* X_3)
- KW method (PS matching) yields unbiased estimates across propensity models containing (1) X_1 and X_2 or (2) X_1 and X_3, with or without interaction terms
- Under the true model, the biases of ALP estimates are consistently closer to zero

Application: Research and Development Survey

National Center for Health Statistics' Research and Development Survey (RANDS)

- Ongoing series of surveys conducted by the National Center for Health Statistics (NCHS, https://www.cdc.gov/nchs/rands/)
- Primarily recruited, web-based commercial survey panels
- Designed to expand NCHS' methodological research:
  - To supplement NCHS' survey and questionnaire evaluation efforts, including the detection of measurement error
  - To explore ways to integrate data from high-quality data collections with commercial survey panels to produce timely estimates while maintaining reliability
- Adapted to provide early experimental estimates on the COVID-19 pandemic (https://www.cdc.gov/nchs/covid19/rands.htm)

Estimating national prevalence of asthma from RANDS 3

- RANDS 3 was conducted in 2019 using NORC's AmeriSpeak Panel
- Panelists were surveyed via web on questions related to general and mental health and medical conditions, including diagnosed asthma
- 2019 National Health Interview Survey, a cross-sectional household interview survey that collects information on a broad range of health topics, is evaluated as the gold standard

		RANDS 3	2019 NHIS
Sample Size		2,646	31,997
Response Rate		18.1%	59.1%
Asthma prevalence	Mean	16.86%	13.46%
	Standard Error	0.98%	0.25%

PS model set up and covariate selection

- Common covariates in RANDS 3 and 2019 NHIS considered as potential calibration variables, including sociodemographic and health variables
- Covariate types were identified using backward selection on outcome and propensity score models containing main effects and pairwise interactions
  - Confounders: common terms in the outcome and propensity models
  - Selection variables and predictors: variables in the propensity model or outcome model only
- All bivariate correlations between selected variables were statistically significant
- PS-adjustment implemented to construct RANDS 3 pseudo-weights using KW method with 2019 NHIS as reference dataset

Estimated asthma prevalence

Propensity Model	Coefficient of Variation	Relative Bias (%)	Standard Error ($\times 10^2$)	Mean Squared Error ($\times 10^4$)
RANDS Weights	0.91	25.31	0.98	12.56
All Variables	1.13	17.55	1.21	7.04
w(x12.n)	1.07	11.41	0.93	3.23
w(x12.r)	1.08	12.85	0.97	3.94
w(x13)	1.10	13.35	1.04	4.31

- Estimate using RANDS weights compared to propensity adjusted weights with variables identified as confounders (x1), predictors from either the NHIS (x2.n) or RANDS (x2.r), and selection variables (x3)
- PS-adjusted estimates had smaller relative bias and MSE compared to estimate using unadjusted RANDS panel weights
- Models containing selection variables produced larger estimated variances

Discussion

Discussion

- Study integrates multiple data sources to provide more robust and efficient inference from web surveys
- Findings provide a principled approach for selecting covariates for population mean estimation
  - Confounders, variables related to both the selection indicator and the outcome of interest, are important to include in the PS model
  - When correlation exists between covariates, the PS model should balance the distributions of the confounder and either the outcome predictor or selection variable
  - The inclusion of selection variables in the PS model will inflate the estimated variance of the population mean but not add bias

References

- Li Y, Irimata K, He Y, Parker J. Variable inclusion strategies through directed acyclic graphs to adjust health surveys subject to selection bias for producing national estimates. Journal of Official Statistics. Accepted.
- Pearl, J (2009). Causality: Models, Reasoning, and Inference. Cambridge, England: Cambridge University Press, 2nd edn.
- 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. https://doi.org/10.1093/biomet/70.1.41.
- Wang, L., Graubard, B.I., Hormuzd, A.K. and Li, Y. 2020. “Improving External Validity of Epidemiologic Cohort Analyses: a Kernel Weighting Approach.” Journal of the Royal Statistical Society Series A, 183, 1293-1311. https://doi.org/10.1111/rssa.12564.

Katherine Irimata
kirimata@cdc.gov

For more information, contact CDC
1-800-CDC-INFO (232-4636)
TTY: 1-888-232-6348 www.cdc.gov

The findings and conclusions in this report are those of the authors and do not necessarily represent the official position of the Centers for Disease Control and Prevention.