Catalogue no. 12-001-X
ISSN 1492-0921

Survey Methodology

Exchangeability assumption in propensity-score based adjustment methods for population mean estimation using non-probability samples

by Yan Li

Release date: June 25, 2024

Statistics Canada Statistique Canada

Canada

How to obtain more information

For information about this product or the wide range of services and data available from Statistics Canada, visit our website, www.statcan.gc.ca.

You can also contact us by

Email at infostats@statcan.gc.ca

Telephone, from Monday to Friday, 8:30 a.m. to 4:30 p.m., at the following numbers:

- Statistical Information Service 1-800-263-1136
- National telecommunications device for the hearing impaired 1-800-363-7629
- Fax line 1-514-283-9350

Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner. To this end, Statistics Canada has developed standards of service that its employees observe. To obtain a copy of these service standards, please contact Statistics Canada toll-free at 1-800-263-1136. The service standards are also published on www.statcan.gc.ca under "Contact us" > "Standards of service to the public."

Note of appreciation

Canada owes the success of its statistical system to a long-standing partnership between Statistics Canada, the citizens of Canada, its businesses, governments and other institutions. Accurate and timely statistical information could not be produced without their continued co-operation and goodwill.

Published by authority of the Minister responsible for Statistics Canada

An HTML version is also available.

Cette publication est aussi disponible en français.

Survey Methodology, June 2024
Vol. 50, No. 1, pp. 37-55
Statistics Canada, Catalogue No. 12-001-X

Exchangeability assumption in propensity-score based adjustment methods for population mean estimation using non-probability samples

Yan Li[1]

Abstract

Nonprobability samples emerge rapidly to address time-sensitive priority topics in different areas. These data are timely but subject to selection bias. To reduce selection bias, there has been wide literature in survey research investigating the use of propensity-score (PS) adjustment methods to improve the population representativeness of nonprobability samples, using probability-based survey samples as external references. Conditional exchangeability (CE) assumption is one of the key assumptions required by PS-based adjustment methods. In this paper, I first explore the validity of the CE assumption conditional on various balancing score estimates that are used in existing PS-based adjustment methods. An adaptive balancing score is proposed for unbiased estimation of population means. The population mean estimators under the three CE assumptions are evaluated via Monte Carlo simulation studies and illustrated using the NIH SARS-CoV-2 seroprevalence study to estimate the proportion of U.S. adults with COVID-19 antibodies from April 01 – August 04, 2020.

Key Words: Balancing score; Propensity-score matching; Propensity-score weighting; Quota sample; Taylor linearization variance estimator; SARS-CoV-2 seroprevalence study.

1. Introduction

Nonprobability samples have emerged rapidly to address time-sensitive priority topics in different areas (Baker, Brick, Bates, Battaglia, Couper, Dever, Gile and Tourangeau, 2013; Kennedy, Mercer, Keeter, Hatley, McGeeney and Gimenez, 2016). These data are timely but subject to selection bias. Participants are often self-selected and volunteer to participate without preassigned selection probabilities. Examples include epidemiological samples that consist of volunteers who are not randomly selected and therefore generally are not representative of any population. Furthermore, volunteers often are subject to "healthy volunteer effects" (Pinsky, Miller, Kramer, Church, Reding, Prorok, Gelmann, Schoen, Buys, Hayes and Berg, 2007), usually resulting in lower estimates of disease incidence and mortality in the volunteer sample than in the general population. Another example is data collected from probability-sampled web panels, which can result in high attrition and nonresponse rates are often found to be 90% or higher (Baker et al., 2013). Although high nonresponse is not necessarily indicative of response bias (Groves and Peytcheva, 2008; Brick and Tourangeau, 2017), selection bias has been of great concern because the composition of web panels often differs from that of the underlying population.

In contrast to nonprobability samples, population-based probability surveys are designed to generate nearly unbiased estimates of population characteristics. They employ probability sample designs, such as stratified multi-stage cluster sampling, to select samples. The resulting samples, when appropriately

1. Yan Li, Joint Program in Survey Methodology and Department of Epidemiology and Biostatistics, University of Maryland at College Park. E-mail: yli6@umd.edu.

weighted by the survey weights, can closely represent the target population and therefore are less susceptible to selection bias.

To reduce selection bias of nonprobability samples, there has been wide literature in survey research investigating the use of propensity-score (PS)-based adjustment methods to improve the population representativeness of nonprobability samples, using probability-based survey samples as external references (Elliott and Valliant, 2017). Various PS-based adjustment methods have been developed and can be grouped into two categories: 1) inverse PS weighting (e.g., Chen, Li and Wu, 2020; Elliott, 2013; Valliant and Dever, 2011) or inverse odds weighting (e.g., Wang, Valliant and Li, 2021) methods (PS-*weighting*); and 2) PS or log-odds of PS matching methods (PS-*matching*) (e.g., Lee and Valliant, 2009; Wang, Graubard, Katki and Li, 2022; Rivers, 2007).

PS-weighting methods construct a pseudoweight for each nonprobability sample unit as the inverse of its propensity of participation. The PS-weighting method corrects for selection bias under the true propensity models, while it can be sensitive to the misspecification of the propensity model (Valliant, 2020) and produce estimates with large variances due to extreme weights (Stuart, 2010). In contrast, the PS-matching method uses the propensity score as a measure of similarity in the distributions of covariates that are included in the propensity model between the probability survey and the nonprobability sample, and thus tend to be less sensitive to the propensity model misspecifications. In addition, the PS-matching method avoids extreme weights, and therefore yields estimates with smaller variances. For recent comprehensive review papers on alternative methods for nonprobability sample analysis and data integration see the papers (Beaumont, 2020; Rao, 2021; Valliant, 2020).

PS-based adjustment methods (e.g., Chen et al., 2020) require the following key assumptions to make nonprobability sample inferences. *First*, the reference survey sample, through weighting, properly represents the finite population (FP) of interest. *Second*, all FP units have a positive participation propensity (i.e., everyone in the population has a positive propensity to participate in nonprobability samples). *Third*, conditional exchangeability (CE) holds without unmeasured covariates, that is, the probability for everyone in the FP to participate in the nonprobability sample is not related to his/her outcome, after conditional on all measured covariates. *Fourth*, being sampled into the reference survey and participating in nonprobability sample are independent. All these assumptions are critical. In this paper, we focus on the CE assumption and examine various balance scores (i.e., functions of covariates) that satisfy the CE assumption.

In observational studies for causal inferences, researchers typically attempt to adjust for all measured covariates, to mimic a completely randomized experiment and assume that such adjustments are sufficient for unbiased estimates of the treatment effects. This assumption is known as "exchangeability of treatment assignment" (Rubin, 1978). In survey research, however, the aim is to make inference about FP parameters and there is little research on the sufficiency of the above-mentioned assumption. Some studies (e.g., Wang et al., 2021) mentioned the need to make assumptions about participation propensity being ignorable given a set of adjustment variables, but aside from noting the presence or absence of biased estimates, there is rarely any additional exploration into whether and to what extent the CE assumption is violated when we make inference about the FP parameters.

The contribution of this paper is to 1) explore the validity of the CE assumption that is conditional on various balancing score estimates that are used in existing PS-based adjustment methods, including both PS-weighting and PS-matching methods, for nonprobability sample inferences; and 2) develop an adaptive balancing score for the CE assumption to improve the efficiency. In this paper, we are not developing new PS-based adjustment methods but study various balancing scores that satisfy the CE assumption. The PS-weighting ALP method is used for illustration purposes. The developed balancing score can also be used in PS-matching methods such as Kernel smoothing method (Wang et al., 2022). The ALP estimators, assuming exchangeability of the outcome conditional on various balancing scores, are evaluated via Monte Carlo simulation studies and illustrated using the NIH SARS-CoV-2 seroprevalence study to estimate the proportion of U.S. adults with COVID-19 antibodies from April 01 – August 04, 2020.

2. Conditional exchangeability (CE) assumption

2.1 Notation

Consider a target finite population (FP) as a random sample of N individuals from a superpopulation model, indexed by $U = \{1, 2, \ldots, N\}$, with observations on a study variable y and a vector of covariates $\mathbf{x}$. Let $\{y_i, \mathbf{x}_i : i \in C\}$ be the observations in the nonprobability sample of individuals, where $C \subset U$ with size n_c. We are interested in estimating the FP mean $\bar{Y}_N = \frac{1}{N}\sum_{i \in U} y_i$ using the nonprobability sample C. The challenge is that we observe C, which, however, may not be a representative sample from U. As a result, $E_C(y \mid U) \neq \bar{Y}_N$, where the subscript C refers to the randomness due to unknown nonprobability sample participation process from U. Let $E(y \mid C) = E_U(E_C(y \mid U))$ and $E(y \mid U) = E_U(\bar{Y}_N)$, where the subscript U refers to the expectation with respect to the superpopulation model. The expectation of y in C may differ from that in U, that is, $E(y \mid C) \neq E(y \mid U)$ due to the selection bias of the nonprobability sample C.

2.2 CE assumption and balancing score

To obtain a design-consistent estimator of $\bar{Y}_N$ using C, the CE assumes

$$E\{y \mid b(\mathbf{x}), C\} = E\{y \mid b(\mathbf{x}), U\}, \tag{2.1}$$

where $b(\mathbf{x})$ is a function of covariates $\mathbf{x}$, so-called balancing score.

The CE assumption (2.1) states that conditional on the balancing score $b(\mathbf{x})$, i.e., a function of measured covariates, the outcome has the same expectation in C as in U. In other words, the nonprobability sample units who carry the same value of balancing score $b(\mathbf{x})$ would represent the same number of FP units. Intuitively, if two persons have the same participation propensity, they would represent the same number of FP units. Therefore, a natural choice of $b(\mathbf{x})$ is the participation propensity $P(i \in C \mid \mathbf{x}, U)$, i.e., the probability for the FP unit i participating in C conditional on the value of $\mathbf{x}$.

More generally, *the basic criterion* for choosing a balancing score is that $b(\mathbf{x})$ is finer than, if not equal to $P(i \in C \mid \mathbf{x}, U)$, to assure the validity the CE assumption (2.1). Therefore, the finest choice of balancing score is $b(\mathbf{x}) = \mathbf{x}$ and the coarsest choice is $b(\mathbf{x}) = P(i \in C \mid \mathbf{x}, U)$ or its monotone function. As a result, the chosen $b(\mathbf{x})$ should be able to distinguish C units with different participation propensities.

In causal inference (Rosenbaum and Rubin, 1983), the conditional exchangeability assumption states that the outcome is exchangeable between the *treated* group vs the *control* group, conditional on all measured covariates. The distribution of covariates in treated group is matched to that in the control group via PS-weighting or PS-matching methods, under a *treatment assignment propensity* model. Treatment effect is then estimated by comparing the two group means after weighting or matching. Analogously, in nonprobability sample inferences, the covariate distribution in the nonprobability sample is matched to that in the FP under a *(nonprobability sample) participation propensity* model. Instead of estimating treatment effect, FP mean is estimated assuming the exchangeability of the outcome between the nonprobability sample and the FP after PS-weighting or PS-matching. Interested readers please refer to Mercer, Kreuter, Keeter and Stuart (2017) for details regarding parallels between causal inference and nonprobability sample inference.

3. Existing balancing scores

3.1 Estimation of $P(i \in C \mid \mathbf{x}, U)$

One can directly estimate the participation propensity $P(i \in C \mid \mathbf{x}, U)$ if covariates $\mathbf{x}$ are known for all individuals in U. Unfortunately, we don't have $\mathbf{x}$ measured for the entire U, whose distribution, however, can be estimated from a probability sample S of size n_s, $\{\mathbf{x}_i : i \in S\}$. Using S as the reference survey, different propensity modeling approaches have been proposed (Chen et al., 2020; Kern, Li and Wang, 2021). For illustration purposes, we assume a logistic regression model

$$\log\left\{\frac{p(\mathbf{x}_i)}{1-p(\mathbf{x}_i)}\right\} = \mathbf{B}^T g(\mathbf{x}_i), \quad \text{for } i \in U, \tag{3.1}$$

where the propensity score $p(\mathbf{x}_i)$ is the propensity of unit i being in the nonprobability sample versus the finite population as approximated by the weighted survey sample, denoted by S_w. Equivalently, $\frac{p(\mathbf{x}_i)}{1-p(\mathbf{x}_i)} = P(i \in C \mid \mathbf{x}_i, U)$. The $g(\mathbf{x}_i)$ is a known function of observed covariates, and $\mathbf{B}$ the unknown regression coefficients to be estimated; see Wang et al. (2021, Section 2.3) for justification of propensity model (3.1). Define w_i is the sample weight of unit $i \in S$. Solving $S(\mathbf{B}) = \left\{\sum_{i' \in C}(1-p(\mathbf{x}_{i'}))\, g(\mathbf{x}_{i'}) - \sum_{i \in S} w_i p(\mathbf{x}_i)\, g(\mathbf{x}_i)\right\} = 0$ for $\mathbf{B}$, the estimate is denoted as $\hat{\mathbf{B}}_w$. The subscript w indicates that the reference survey weights are used to estimate $\mathbf{B}$ in the propensity model (3.1). The participation propensity $P(i \in C \mid \mathbf{x}, U)$ for $i \in C \cup S$ can be estimated by $\exp(\mathbf{x}_i \hat{\mathbf{B}}_w) = \frac{\hat{p}(\mathbf{x}_i, \hat{\mathbf{B}}_w)}{1-\hat{p}(\mathbf{x}_i, \hat{\mathbf{B}}_w)}$, with $\hat{p}(\mathbf{x}_i, \hat{\mathbf{B}}_w)$ the estimate of the propensity score $p(\mathbf{x}_i)$.

3.2 CE assumption conditional on $b(\mathbf{x};\hat{\mathbf{B}}_w)$

To satisfy the CE assumption (2.1), the balancing score should be, as fine as or finer than, the estimated participation rate. Following Wang et al. (2022), the linear predictor, i.e., a natural log transformation of the estimated participation propensity, is used as the balance score, i.e., $b(\mathbf{x}_i;\hat{\mathbf{B}}_w)=\hat{\mathbf{B}}_w^T g(\mathbf{x}_i)=\log \hat{p}(i\in C|\mathbf{x}_i,U)$. Therefore, under the propensity model (3.1), let $b(\mathbf{x})=b(\mathbf{x};\hat{\mathbf{B}}_w)$ in (2.1), that is,

$$E\{y\,|\,b(\mathbf{x};\hat{\mathbf{B}}_w),C\}=E\{y\,|\,b(\mathbf{x};\hat{\mathbf{B}}_w),U\}$$

approximately holds. As follows, we estimate population mean by various existing PS-based adjustment methods. For example, the PS-weighting method ALP (Wang et al., 2021) weights the unit i in C by inverse of $\hat{p}(i\in C\,|\,\mathbf{x}_i,U)=\exp(b(\mathbf{x}_i;\hat{\mathbf{B}}_w))$. Another example is the PS-matching method KW (Wang et al. 2022), which matches the units in C and S based on the similarity in $b(\mathbf{x};\hat{\mathbf{B}}_w)$. It has been proved that both the ALP and the KW estimates are approximately unbiased under the CE assumption conditional on $b(\mathbf{x};\hat{\mathbf{B}}_w)$.

A severe drawback of this method, however, is the potentially large variance inflation in $b(\mathbf{x};\hat{\mathbf{B}}_w)$ due to the variability (Scott and Wild, 2001; Li, Graubard and DiGaetano, 2011) of the differential reference survey weights *versus* the nonprobability sample weights (= 1) in estimating the model parameter $\mathbf{B}$. For variance reduction, the survey weights have been ignored in estimating $\mathbf{B}$ (Wang, Graubard, Katki and Li, 2020; Lee and Valliant, 2009).

3.3 CE assumption conditional on $b(\mathbf{x};\hat{\mathbf{B}}_0)$

Assume the propensity model (3.1) is fitted to the combined nonprobability sample and unweighted survey data $(C\cup S)$ and the resulting estimated balancing score is $b(\mathbf{x}_i,\hat{\mathbf{B}}_0)=\hat{\mathbf{B}}_0^T g(\mathbf{x}_i)$ with $\hat{\mathbf{B}}_0$ obtained by solving the estimating equation $S(\mathbf{B})=\left\{\sum_{i'\in C}(1-p(\mathbf{x}_{i'}))\,g(\mathbf{x}_{i'})-\sum_{i\in S}p(\mathbf{x}_i)\,g(\mathbf{x}_i)\right\}=0$ for $\mathbf{B}$, without considering probability sample weights. Accordingly, the CE based on $\hat{\mathbf{B}}_0$, assumed by existing PS-weighting or PS-matching methods (Wang et al., 2020; Lee and Valliant, 2009; Kern et al., 2021), is

$$E\{y\,|\,b(\mathbf{x};\hat{\mathbf{B}}_0),C\}=E\{y\,|\,b(\mathbf{x};\hat{\mathbf{B}}_0),U\}.$$

Without using the survey weights, the estimated balancing score $b(\mathbf{x},\hat{\mathbf{B}}_0)$ can be more stable than the $b(\mathbf{x},\hat{\mathbf{B}}_w)$. The question is how plausible is the CE assumption conditional on $b(\mathbf{x},\hat{\mathbf{B}}_0)$ in real problems.

Note $b(\mathbf{x},\hat{\mathbf{B}}_0)$ produces balanced $\mathbf{x}$ distribution between C and S, and therefore exchangeability of the y distribution (with all $\mathbf{x}$ balanced) holds between C and S, which, however, is not sufficient to obtain an unbiased estimate of the FP mean. Instead, the exchangeability of y distribution between C and U conditional on $b(\mathbf{x},\hat{\mathbf{B}}_0)$ is required. We know from Section 2.2 that $P(i\in C\,|\,\mathbf{x},U)$ is the coarsest balancing score that satisfies (2.1) and $b(\mathbf{x},\hat{\mathbf{B}}_w)$ approximately produce a balanced y distribution between C and U. According to the basic criteria for choosing balancing score, the $b(\mathbf{x},\hat{\mathbf{B}}_0)$ needs to be as fine as or finer than $b(\mathbf{x},\hat{\mathbf{B}}_w)$ One example is that $b(\mathbf{x}_i;\hat{\mathbf{B}}_0)=\hat{\mathbf{B}}_0^T g(\mathbf{x}_i)$ is a linear function of $b(\mathbf{x}_i;\hat{\mathbf{B}}_w)=\hat{\mathbf{B}}_w^T g(\mathbf{x}_i)$, that is,

$\hat{\mathbf{B}}_w^T = \text{const.} \times \hat{\mathbf{B}}_0^T$. Suppose the reference survey S oversamples, by design, a minority group, say, African American females. This linear relationship requires that the distribution of the same minority group, defined by race/ethnicity and gender, in the nonprobability sample should be proportional to that in the reference survey. In reality, however, we have no control over the nonprobability sampling, and therefore the linear relationship may hold only by chance. The estimator based on $b(\mathbf{x}; \hat{\mathbf{B}}_0)$ is efficient but can be biased.

A hypothetical example

For illustration purpose, suppose a nonprobability sample and a survey sample are selected by probability proportional to size (PPS) sampling with the measure of size for FP unit i, defined by, respectively, $s_{ic} = \exp(x_{i1}B_1 + x_{i2}B_2)$ for the nonprobability sample participation and $s_{is} = \exp(x_{i1}B_1' + x_{i3}B_3)$ for the survey sample selection. Suppress the subscript i and let $B_1 \approx B_1'$. The probability of a FP unit participating in the nonprobability sample (p_c) *versus* being selected into the survey (p_s) is

$$\begin{aligned}\log\left(\frac{p_c}{p_s}\right) &= \log\left(\frac{n_c s_c}{\sum_U s_c} \Big/ \frac{n_s s_s}{\sum_U s_s}\right) = \log\left(\frac{n_c \sum_U s_s}{n_s \sum_U s_c} \times \frac{s_c}{s_s}\right), \\ &= \text{const.} + x_1(B_1 - B_1') + x_2 B_2 - x_3 B_3 = \text{const.} + \mathbf{x}^T \mathbf{B}_0,\end{aligned}$$

where $\mathbf{x} = (x_1, x_2, x_3)^T$ and $\mathbf{B}_0 = (B_1 - B_1', B_2, -B_3)^T$. By fitting a logistic model, including all variables $\mathbf{x}$, to the combined (nonprobability and unweighted survey) sample, an estimated balancing score would be

$$b(\mathbf{x}; \hat{\mathbf{B}}_0) = \mathbf{x}^T \hat{\mathbf{B}}_0.$$

Note $\hat{\mathbf{B}}_0$ includes the attenuated x_1 effect in constructing $b(\mathbf{x}; \hat{\mathbf{B}}_0)$ due to the similar x_1 distribution in S and in C. As a result, the estimated balancing scores cannot distinguish the C units with different participation propensities by x_1, and thus $E\{y \mid b(\mathbf{x}; \hat{\mathbf{B}}_0), C\} \neq E\{y \mid b(\mathbf{x}; \hat{\mathbf{B}}_0), U\}$, leading to biased estimation of $\bar{Y}_N$.

In next section, we propose an adaptive balancing score that adjusts $b(\mathbf{x}; \hat{\mathbf{B}}_0)$ to be a monotone function of the estimated $P(i \in C \mid \mathbf{x}, U)$ for unbiased estimation of the FP mean.

4. Adaptive balancing score

We propose an adjusted balancing score in three steps. The first step is to fit a logistic regression model to the combined $C \cup S$ sample without weights, given by (Wang et al., 2020)

$$\log\left\{\frac{p(i \in C \mid \mathbf{x}_i, U)}{p(i \in S \mid \mathbf{x}_i, U)}\right\} = \log\left\{\frac{p^*(\mathbf{x}_i)}{1 - p^*(\mathbf{x}_i)}\right\} = \mathbf{B}_0^T g(\mathbf{x}_i) \quad \text{for } i \in U \tag{4.1}$$

and the estimates of the model parameter $\mathbf{B}_0$ are denoted as $\hat{\mathbf{B}}_0$, where $p^*(\mathbf{x}_i)$ is the propensity of being in C vs. in S for unit i. As discussed in Section 3.3, $b(\mathbf{x}; \hat{\mathbf{B}}_0) = \hat{\mathbf{B}}_0^T g(\mathbf{x}_i)$ balances the $\mathbf{x}$ distribution

between C and S. Without considering sample weights in the analysis, $\hat{\mathbf{B}}_0$ tends to be more efficient than $\hat{\mathbf{B}}_w$. The CE assumption (2.1), however, can be violated when $b(\mathbf{x}; \hat{\mathbf{B}}_0)$ fails at balancing the distribution in $\mathbf{x}$ between C and U.

The second step aims to develop a bias correction factor to adjust $b(\mathbf{x}; \hat{\mathbf{B}}_0)$ so that the balanced distribution in $\mathbf{x}$ between C and U (approximated by the weighted reference survey S_w) can be achieved. As a *computational* device, a pseudo-population of $S^* \cup U$ is constructed, where S^* is a duplicate of S that has the same joint distributions of covariates $\mathbf{x}$ and outcome y as the original S. In the pseudo-population $S^* \cup U$, S^* and S are treated as two different sets. We model $q(\mathbf{x}_i)$ as the probability for unit i to be included in S from the pseudo-population, that is,

$$q(\mathbf{x}_i) = p(i \in S \mid \mathbf{x}_i, S^* \cup U) = \frac{p(i \in S \mid \mathbf{x}_i, U)}{1 + p(i \in S \mid \mathbf{x}_i, U)}.$$

Assume a logistic model

$$\log\{p(i \in S \mid \mathbf{x}_i, U)\} = \log\left\{\frac{q(\mathbf{x}_i)}{1 - q(\mathbf{x}_i)}\right\} = \boldsymbol{\gamma}^T g(\mathbf{x}_i), \quad \text{for } i \in U \tag{4.2}$$

where $\boldsymbol{\gamma}$ denotes the model parameters, estimated by solving the estimating equation $S(\boldsymbol{\gamma}) = \sum_{i \in S}(1 - q(\mathbf{x}_i) - w_i q(\mathbf{x}_i))\, g(\mathbf{x}_i) = 0$ for $\boldsymbol{\gamma}$. The estimate is denoted by $\hat{\boldsymbol{\gamma}}_w$, measuring the effects of $g(\mathbf{x})$ on the sample S selection. We use it for the correction of distorted or missing effects of $g(\mathbf{x})$ on the nonprobability sample C participation propensity in $b(\mathbf{x}; \hat{\mathbf{B}}_0)$, especially for those variables involved in both S sampling and C participating processes.

At step 3, the new balancing score estimate is constructed as

$$b(\mathbf{x}_i; \hat{\mathbf{B}}_0, \hat{\boldsymbol{\gamma}}_w) = (\hat{\boldsymbol{\gamma}}_w^T + \hat{\mathbf{B}}_0^T)\, g(\mathbf{x}_i) \quad \text{for } i \in U.$$

As noted, adding up models (4.1) and (4.2) yields model (3.1), with the left side equal to

$$\log\left\{\frac{p(i \in C \mid \mathbf{x}_i, U)}{p(i \in S \mid \mathbf{x}_i, U)}\right\} + \log\{p(i \in S \mid \mathbf{x}_i, U)\} = \log\{P(i \in C \mid \mathbf{x}_i, U)\},$$

a monotone function of participation propensity, and the right side the same functional form $g(\mathbf{x}_i)$ as in model (3.1). We know that $b(\mathbf{x}; \hat{\mathbf{B}}_w)$ under model (3.1), although satisfying the CE assumption (2.1), can be inefficient due to differential weights in the analysis. Instead of fitting the model (3.1) directly to the combined nonprobability and weighted survey data $(C \cup S_w)$ to obtain $\hat{\mathbf{B}}_w$, we construct the adjusted balancing score $b(\mathbf{x}; \hat{\mathbf{B}}_0, \hat{\boldsymbol{\gamma}}_w)$ based on $\hat{\mathbf{B}}_0$ and $\hat{\boldsymbol{\gamma}}_w$ in three steps. This adjusted balancing score is a monotone (natural logarithm) function of sample C participation propensity, and therefore the y distribution is exchangeable between C and U, that is,

$$E\{y \mid b(\mathbf{x}; \hat{\mathbf{B}}_0, \hat{\boldsymbol{\gamma}}_w^T), C\} = E\{y \mid b(\mathbf{x}; \hat{\mathbf{B}}_0, \hat{\boldsymbol{\gamma}}_w^T), U\}$$

approximately holds.

As follows, PS-based adjustment methods can be conducted to create pseudoweights for units in C based on the new adaptive balancing score $b(\mathbf{x}; \hat{\mathbf{B}}_0, \hat{\boldsymbol{\gamma}}_w^T)$. PS-weighting methods weight each unit in C by the inverse of estimated participation rate. In contrast, PS-matching methods match C and S units based on the adaptive balancing score and then distributes the sample weights in S to C units according to their similarities. For example, ALP weighting method (Wang et al., 2021) creates pseudoweights

$$\hat{w}_j^{\text{ALP}} = \exp^{-1}(b(\mathbf{x}_j; \hat{\mathbf{B}}_0, \hat{\boldsymbol{\gamma}}_w^T)) \quad \text{for } j \in C.$$

Kernel smoothing method (Wang et al., 2022) creates pseudoweights by adding up the fractional weights distributed from each survey unit $i \in S$,

$$\hat{w}_j^{\text{KW}} = \sum_{i \in S} w_i K_{ij} \quad \text{with} \quad K_{ij} = \frac{K\left(\frac{d_{ij}}{h}\right)}{\sum_{l \in C} K\left(\frac{d_{il}}{h}\right)} \quad \text{for } j \in C,$$

where $K(\cdot)$ is an arbitrary kernel function such as standard normal density function, h is the bandwidth associated with $K(\cdot)$, and the distance $d_{ij} = b(\mathbf{x}_i; \hat{\mathbf{B}}_0, \hat{\boldsymbol{\gamma}}_w^T) - b(\mathbf{x}_j; \hat{\mathbf{B}}_0, \hat{\boldsymbol{\gamma}}_w^T)$ measures the similarity in $\mathbf{x}$ distribution between the nonprobability sample unit $j \in C$ and the survey unit $i \in S$.

The population mean can then be estimated by

$$\bar{y} = \frac{1}{\sum_{j \in C} \hat{w}_j} \sum_{j \in C} \hat{w}_j y_j, \tag{4.3}$$

where $\hat{w}_j$ can be $\hat{w}_j^{\text{ALP}}$ or $\hat{w}_j^{\text{KW}}$.

To estimate the variance of $\bar{y}$, we assume the FP size $N \to \infty$ and consider the randomness due to sampling of S and participation process of C from U. Taylor linearization (TL) variance estimator is developed to account for the variability due to estimating propensity scores $p^*(\mathbf{x}_i)$ and $q(\mathbf{x}_i)$ in steps 1-2. The TL technique is commonly employed in the survey literature to derive design-consistent variance estimators (Li, Graubard, Huang and Gastwirth, 2015; Li and Graubard, 2012). Assuming independence between being sampled to the reference survey and participating in the nonprobability sample, the variance of $\bar{y}$ can be approximated by (Korn and Graubard, 1999)

$$\text{var}_{\text{TL}}(\bar{y}) \cong \text{var}\left(\sum_{j \in C} z_j\right) + \text{var}\left(\sum_{i \in S} z_i\right), \tag{4.4}$$

where z_j (or z_i) is the Taylor deviate (TD) for j^{th} (or i^{th}) unit in C (or in S) derived by taking the derivative of $\bar{y}$ with respect to the sample weight (Shah, 2004). For example, when $\hat{w}_j = \hat{w}_j^{\text{ALP}}$, the TD for unit $j \in C$ is

$$z_j = \frac{\partial}{\partial w_j} \bar{y} = \frac{\hat{w}_j (y_j - \bar{y})}{\sum_{l \in C} \hat{w}_l} + \frac{\sum_{l \in C} (y_l - \bar{y})}{\sum_{l \in C} \hat{w}_l} \left(\frac{\partial}{\partial w_j} \hat{w}_l\right)$$

and

$$\frac{\partial}{\partial w_j}\hat{w}_l = \left(\frac{\partial}{\partial \hat{\theta}}\hat{w}_l\right)\left(\frac{\partial}{\partial w_j}\hat{\theta}\right) = -\hat{w}_l x_l \left(\frac{\partial}{\partial w_j}\hat{\theta}\right),$$

where $\hat{\theta}$ is the estimated model parameters, which can be $\hat{B}_0$, $\hat{B}_w$, or $\hat{B}_0 + \hat{\gamma}_w$, e.g.,

$$\frac{\partial}{\partial w_j}(\hat{B}_0 + \hat{\gamma}_w) = (1-\hat{p}_j^*)\, x_j \left(\sum_{j' \in C \cup S} \hat{p}_{j'}^*(1-\hat{p}_{j'}^*)\, x_{j'} x_{j'}^T\right)^{-1},$$

where $\hat{p}_j^*$ for $j \in C$ is the estimated propensity score for unit j under model (4.1).

For unit $i \in S$, the TD is

$$z_i = \frac{\sum_{j \in C}(y_j - \bar{y})}{\sum_{j \in C}\hat{w}_j}\left(\frac{\partial}{\partial w_i}\hat{w}_j\right),$$

and

$$\frac{\partial}{\partial w_i}\hat{w}_j = \frac{\partial}{\partial \hat{\theta}}\hat{w}_j \frac{\partial}{\partial w_i}\hat{\theta} = -\hat{w}_j x_j \left(\frac{\partial}{\partial w_i}\hat{\theta}\right),$$

where $\hat{\theta}$ can be $\hat{B}_0$, $\hat{B}_w$, or $\hat{B}_0 + \hat{\gamma}_w$ e.g.,

$$\begin{aligned}\frac{\partial}{\partial w_i}(\hat{B}_0 + \hat{\gamma}_w) \quad &= -\hat{p}_i^* x_i \left(\sum_{j' \in C \cup S} \hat{p}_{j'}^*(1-\hat{p}_{j'}^*)\, x_{j'} x_{j'}^T\right)^{-1} \\ &\quad + \left(1 - \hat{q}_i - \hat{q}_i w_i\right) x_i \left(\sum_{j' \in S}(1+w_{j'})\,\hat{q}_{j'}\,(1-\hat{q}_{j'})\, x_{j'} x_{j'}^T\right)^{-1},\end{aligned}$$

where $\hat{q}_i$ for $i \in S$ is the estimated propensity score for unit i under model (4.2). TD for each unit measures the change of the nonlinear estimator, in our case $\bar{y}$, as if the unit was deleted from the sample. TL variance estimator of $\bar{y}$ is then approximated by (4.4), where $\text{var}\left(\sum_{i \in S} z_i\right)$ accounts for variability due to complex sampling of S. Following Wang et al. (2021), it can be proved that $\bar{y}$ is design-consistent and $\text{var}_{\text{TL}}(\bar{y}) = O\left(\frac{1}{n_c}\right) + O\left(\frac{1}{n_s}\right)$. Sections 5 and 6 report the ALP estimates for illustration of the exchangeability assumptions conditional on various balancing scores. Similarly, the variance estimators of KW estimates with the adaptive balance scores can be derived, which will be given in a future paper.

5. Simulation studies

5.1 Population generation

Simulation studies are conducted to evaluate the ALP estimates based on the adjusted balancing score $b(\mathbf{x}; \hat{\mathbf{B}}_0, \hat{\boldsymbol{\gamma}}_w^T)$, along with $b(\mathbf{x}; \hat{\mathbf{B}}_0)$ and $b(\mathbf{x}; \hat{\mathbf{B}}_w)$ for comparison purpose. We generate a finite population

(FP) of size $N = 1{,}000{,}000$ with three independent covariates x_1, x_2, x_3, each following a standard normal distribution $N(0,1)$. The binary outcome $\mathbf{Y}$ is generated with mean defined by

$$P(Y=1) = \frac{\exp(\beta_0 + x_1\beta_{x_1} + x_2\beta_{x_2} + x_1x_2\beta_{x_1x_2})}{1+\exp(\beta_0 + x_1\beta_{x_1} + x_2\beta_{x_2} + x_1x_2\beta_{x_1x_2})}, \tag{5.1}$$

where $\beta_y = (\beta_0, \beta_{x_1}, \beta_{x_2}, \beta_{x_1x_2})^T$ are the outcome model parameters specified as $\beta_0 = -1$, $\beta_{x_1} = 0.8$, $\beta_{x_2} = 0.2$, $\beta_{x_1x_2} = 0.5$. The average of the binary outcome is about 30%. The results showed a similar pattern when $\beta_0 = -2$ or -3 and hence are not shown.

5.2 Probability sample S selection

We select a random probability sample S of size n_s with replacement from the FP using probability proportional to size (PPS) sampling with the measure of size for the k^{th} FP individual (mos_k) defined by

$$\text{mos}_k = \exp\left[a \times (\alpha_0 + x_{k1}\alpha_{x_1} + x_{k2}\alpha_{x_2} + x_{k3}\alpha_{x_3} + + x_{k1}x_{k2}\alpha_{x_1x_2} + x_{k1}x_{k3}\alpha_{x_1x_3})\right] \tag{5.2}$$

so that the inclusion probability is

$$p(k \in S \mid x; U) = \frac{n_s \times \text{mos}_k}{\sum_{k\in U}\text{mos}_k},$$

and the corresponding sample weight is the inverse of the inclusion probability, i.e., $w_k = \frac{\sum_{k\in U}\text{mos}_k}{n_s \times \text{mos}_k}$. We specify $\left(\alpha_0, \alpha_{x_1}, \alpha_{X_2}, \alpha_{x_3}, \alpha_{x_1x_2}, \alpha_{x_1x_3}\right) = (-1,\ 0.5,\ 0,\ 0.5,\ 0,\ -0.2)$ and let $a = 0.5$, 1, or 1.5 to vary the coefficient of variation (CV) of the sample weights in S (denoted by w_s), corresponding to $\text{CV}(w_s) =$ 0.38, 0.86, or 1.5, respectively. Note the selection variables in sampling S are x_1 and x_3, and the w_k-weighted probability sample S approximates the FP.

5.3 Nonprobability sample C selection

The underlying selection process for sampling C is unknown. We select C of size $n_c = 2{,}500$ from the FP using PPS sampling with mos_k, given by (5.2) and specified to include three scenarios: 1) quota sample that has the same joint distribution of x_1 and x_2 as in the FP, denoted by Quota.x_1x_2; 2) quota sample that has the same distribution of x_2 as in the FP, denoted by Quota.x_2; and 3) a volunteer sample with different distributions in x_1 or x_2 from those in the FP, denoted by Volunteer. Variable x_3 is not predictive of the outcome and therefore induce no bias in FP mean estimation (Li, Irimata, He and Parker, 2022). Table 5.1 summarizes the model parameters for the outcome generation in (5.1), the probability sample S selection and three nonprobability samples selection in (5.2). We vary the probability sample size $n_s = 1{,}250$; 2,500; 3,750 and the nonprobability sample size is fixed at $n_c = 2{,}500$. Sample weights associated with C units are masked in the analysis.

Table 5.1
Model parameter specifications for outcome generation, probability sample (*S*) selection, and nonprobability sample (*C*) selection.

Model	Intercept	x_1	x_2	x_3	x_1x_2	x_1x_3
Outcome	-1	0.8	0.2	0	0.5	0
Sample S Selection	-1	0.5	0	0.5	0	-0.2
Sample C Participation						
Quota. x_1x_2	-1	0	0	0.5	0	0
Quota. x_2	-1	0.5	0	0	0	-0.2
Volunteer	-1	0.5	0.5	0.5	0	-0.2

The three ALP estimates (4.3) based on the adaptive balancing score $b(\mathbf{x}; \hat{\mathbf{B}}_0, \hat{\boldsymbol{\gamma}}_w^T)$, unweighted $b(\mathbf{x}; \hat{\mathbf{B}}_0)$ and weighted $b(\mathbf{x}; \hat{\mathbf{B}}_w)$ are computed for each of the $R =$ 1,000 simulation runs and evaluated by

- Relative Bias (RelBias%) = Bias (= average of R simulated means $-$ population mean) divided by population mean $\times 100\%$.
- Empirical Variance (EV) = Variance of R simulated means $\times 10^4$.
- Variance Ratio (VR) = TL Variance/Empirical Variance.

To construct the three balancing score estimate, the function $g(\mathbf{x}_i)$ in (3.1)-(4.2) includes not only the main effects of x_1, x_2, x_3 but also their pairwise interaction effects. It is expected that ALP estimates with $b(\mathbf{x}; \hat{\mathbf{B}}_w)$ are approximately unbiased but with inflated variance due to differential weights; ALP estimates with $b(\mathbf{x}; \hat{\mathbf{B}}_0)$ have the smallest variance but can be biased. In contrast, it is expected that the ALP estimates with the adaptive balancing score $b(\mathbf{x}; \hat{\mathbf{B}}_0, \hat{\boldsymbol{\gamma}}_w^T)$ are approximately unbiased with smaller variance under the true propensity models.

5.4 Results

Table 5.2 presents the relative bias (%) of ALP estimates with the three balancing scores using nonprobability samples of Quota. x_1x_2, Quota. x_2, and Volunteer. For comparison purpose, we also include the unweighted estimates. We make three observations: 1) As expected, unweighted estimates are unbiased for Quota. x_1x_2, but badly biased for Quota. x_2 and Volunteer samples. This result is consistent with the findings in Li et al. (2022). 2) As a remedy, the balance scores of $b(\mathbf{x}; \hat{\mathbf{B}}_w)$ or $b(\mathbf{x}; \hat{\mathbf{B}}_0, \hat{\boldsymbol{\gamma}}_w^T)$ match the Quota. x_2 or the Volunteer sample to the joint distribution of x_1 and x_2 in the FP and therefore produce approximately unbiased estimates across all three nonprobability samples. 3) In contrast, the unweighted score $b(\mathbf{x}; \hat{\mathbf{B}}_0)$ leads to biased estimates since it is not a monotone or finer function of the estimated participation propensity of all three nonprobability samples.

Table 5.2
Relative Bias (%) of ALP estimates of population mean ($\bar{Y} = 0.3$) with the probability and nonprobability sample sizes $n_s = n_c$ = 2,500 and CV(w_s) = 0.86.

	Quota. $x_1 x_2$ CV(w_c) = 0.53	Quota. x_2 CV(w_c) = 0.6	Volunteer CV(w_c) = 1.10
Unweighted	-1.33	24.33	33.67
$b(\mathbf{x}; \hat{\mathbf{B}}_w)$	-1.33	-1.33	-1.33
$b(\mathbf{x}; \hat{\mathbf{B}}_0)$	19.33	19.33	19.33
$b(\mathbf{x}; \hat{\mathbf{B}}_0, \hat{\boldsymbol{\gamma}}_w^T)$	-1.33	-1.33	-1.33

Next, we compare in Table 5.3 the two unbiased ALP estimates with $b(\mathbf{x}; \hat{\mathbf{B}}_0, \hat{\boldsymbol{\gamma}}_w^T)$ and $b(\mathbf{x}; \hat{\mathbf{B}}_w)$ in terms of their efficiency when varying the coefficients of variation (CV) of the probability sample weights CV(w_s) = 0.38, 0.86, or 1.50. We make three observations. *First*, as the CV(w_s) increases, the variance increases as expected. For example, when using $b(\mathbf{x}; \hat{\mathbf{B}}_w)$ the empirical variance increases from 1.00 to 1.12 to 1.30 for Quota. $x_1 x_2$. *Second*, as the CV(w_s) increases, the efficiency gain of $b(\mathbf{x}; \hat{\mathbf{B}}_0, \hat{\boldsymbol{\gamma}}_w^T)$ over $b(\mathbf{x}; \hat{\mathbf{B}}_w)$ becomes larger. For example, the relative difference of the two empirical variances increases from 1% (= (1 − 0.99) / 1.00) to 4% (= (1.12 − 1.07) / 1.12) to 12% (= (1.3 − 1.14) / 1.30) when CV(w_s) increases from 0.38 to 0.86 to 1.5 for Quota. $x_1 x_2$. *Third*, comparing the three nonprobability samples, the efficiency gain of $b(\mathbf{x}; \hat{\mathbf{B}}_0, \hat{\boldsymbol{\gamma}}_w^T)$ over $b(\mathbf{x}; \hat{\mathbf{B}}_w)$ is largest for Quota. $x_1 x_2$. Intuitively, the pseudoweights created for Quota. $x_1 x_2$ are noninformative and thus add extra variance due to estimating $b(\mathbf{x}; \hat{\mathbf{B}}_w)$.

Table 5.3
Empirical Variance ($\times 10^4$) of two unbiased ALP estimates by varying coefficients of variation of probability sample weights CV(w_s), $n_s = n_c$ = 2,500.

	Quota. $x_1 x_2$	Quota. x_2	Volunteer
	CV(w_s) = 0.38		
Unweighted	0.81	0.94	0.81
$b(\mathbf{x}; \hat{\mathbf{B}}_w)$	1.00	0.97	1.44
$b(\mathbf{x}; \hat{\mathbf{B}}_0, \hat{\boldsymbol{\gamma}}_w^T)$	0.99	0.98	1.45
	CV(w_s) = 0.86		
Unweighted	0.85	0.90	0.99
$b(\mathbf{x}; \hat{\mathbf{B}}_w)$	1.12	1.00	1.62
$b(\mathbf{x}; \hat{\mathbf{B}}_0, \hat{\boldsymbol{\gamma}}_w^T)$	1.07	1.02	1.64
	CV(w_s) = 1.50		
Unweighted	0.85	0.90	0.99
$b(\mathbf{x}; \hat{\mathbf{B}}_w)$	1.30	1.11	1.72
$b(\mathbf{x}; \hat{\mathbf{B}}_0, \hat{\boldsymbol{\gamma}}_w^T)$	1.14	1.07	1.68

Table 5.4 presents the empirical variance (EV) in the left panel and the variance ratio (VR) in the right panel for the ALP estimates when varying the probability sample sizes (n_s = 1,250; 2,500; 3,750) with a fixed nonprobability sample size n_c = 2,500. We make the following three observations. *First*, EV decreases as n_s increases, e.g., EV of ALP estimates with $b(\mathbf{x}; \hat{B}_w)$ for Quota. $x_1 x_2$ decreases from 1.33 to 1.08 to 0.99. However, the difference becomes smaller, i.e., a larger EV drop of 0.25 (= 1.33 − 1.08), as

n_s increases from 1,250 to 2,500, compared to a moderate drop of 0.09 (= 1.08 − 0.99), as n_s increases from 2,500 to 3,750. This result is because $\text{Var}(\bar{y}) = O\left(\frac{1}{n_c}\right) + O\left(\frac{1}{n_s}\right)$ is dominated by $O\left(\frac{1}{n_c}\right)$ when $n_s > n_c$ and therefore the efficiency gain is moderate by increasing n_s once $n_s > n_c$. *Second,* comparing the two balancing scores, $b(\mathbf{x}; \hat{B}_0, \hat{\gamma}_w^T)$ is more efficient than $b(\mathbf{x}; \hat{B}_w)$ when n_s is small. Intuitively, when $n_s < n_c$, the variance of ALP estimates is dominated by the probability sample S, which has differential sample weights w_s and therefore induces large variability when estimating $\hat{B}_w$. This occurs especially for quota samples where sample weights used in estimating $b(\mathbf{x}; \hat{B}_w)$ are approximately noninformative and thus add extra variance. *Third,* the proposed TL variance estimator generally performs well with variance ratio's close to one (see right panel in Table 5.4). The TL variance with $b(\mathbf{x}; \hat{B}_0, \hat{\gamma}_w^T)$, however, overestimates the variance for Quota. $x_1 x_2$ when n_s is small. It is found that the VR for $b(\mathbf{x}; \hat{B}_0, \hat{\gamma}_w^T)$ is closer to one as n_s increases or when $\text{CV}(w_s)$ is small (results not shown).

Table 5.4
Empirical Variance ($\times 10^4$) and Variance Ratio of two unbiased ALP estimates with varying probability sample sizes n_s, $\text{CV}(w_s) = 0.86$ and $n_c = 2{,}500$.

	Empirical Variance (EV)			Variance Ratio (VR)		
	Quota. $x_1 x_2$	Quota. x_2	Volunteer	Quota. $x_1 x_2$	Quota. x_2	Volunteer
Unweighted	0.81	0.87	1.03	1.02	0.95	0.86
			n_s = 1,250			
$b(\mathbf{x}; \hat{\mathbf{B}}_w)$	1.33	1.18	1.77	1.02	0.98	0.96
$b(\mathbf{x}; \hat{\mathbf{B}}_0, \hat{\boldsymbol{\gamma}}_w^T)$	1.17	1.08	1.80	1.41	1.35	1.14
			n_s = 2,500			
$b(\mathbf{x}; \hat{\mathbf{B}}_w)$	1.08	0.98	1.65	1.06	1.01	0.93
$b(\mathbf{x}; \hat{\mathbf{B}}_0, \hat{\boldsymbol{\gamma}}_w^T)$	1.00	0.96	1.69	1.31	1.23	1.03
			n_s = 3,750			
$b(\mathbf{x}; \hat{\mathbf{B}}_w)$	0.99	0.94	1.60	1.08	1.00	0.93
$b(\mathbf{x}; \hat{\mathbf{B}}_0, \hat{\boldsymbol{\gamma}}_w^T)$	0.95	0.94	1.63	1.26	1.15	1.00

In summary, via simulation studies, it is observed ALP estimates with $b(\mathbf{x}; \hat{\mathbf{B}}_0, \hat{\boldsymbol{\gamma}}_w^T)$ and $b(\mathbf{x}; \hat{\mathbf{B}}_w)$ are approximately unbiased and comparably efficient when the reference probability sample has large sample size n_s or stable sample weights with small $\text{CV}(w_s)$. In contrast, when the reference probability sample has small n_s or variable sample weights, $b(\mathbf{x}; \hat{\mathbf{B}}_0, \hat{\boldsymbol{\gamma}}_w^T)$ tend to produce more efficient estimates, especially for quota samples. Survey practitioners should choose a reference survey with sufficiently large size and stable sample weights, and note that the efficiency gain by increasing n_s is moderate once $n_s > n_c$.

6. NIH SARS-CoV-2 seropositivity data analysis

The primary purpose of the SARS-CoV-2 Seropositivity Study is to estimate the prevalence of seropositivity to the SARS-CoV-2 virus antibody in the target population consisting of adults 18+ years old living in the US who had not been diagnosed with Covid during the early phase of the pandemic from April

to August 2020. Within weeks of the study recruitment announcement, more than 460,000 individuals volunteer. The study, however, could only afford a subset of these volunteers. A quota sample were selected based on six quota variables, age group, race, sex, ethnicity, population density, and geographic region, to approximately match the US adults' distribution of these variables. There were 8,058 participants who responded to the questionnaire about clinical factors and provided blood samples for assessing seropositivity. The sample collected from the SARS-CoV-2 Seropositivity Study will be called "covid sample". Although the covid sample was a random sample with known selection probabilities from the pool of volunteers Kalish, Klumpp-Thomas, Hunsberger, Baus, Fay, Siripong, Wang, Hicks, Mehalko, Travers, Drew, Pauly, Spathies, Ngo, Adusei, Karkanitsa, Croker, Li, Graubard, Czajkowski, Belliveau, Chairez, Snead, Frank, Shunmugavel, Han, Giurgea, Rosas, Bean, Athota, Cervantes-Medina, Gouzoulis, Heffelfinger, Valenti, Caldararo, Kolberg, Kelly, Simon, Shafiq, Wall, Reed, Ford, Lokwani, Denson, Messing, Michael, Gillette, Kimberly, Reis, Hall, Esposito, Memoli and Sadtler (2021), this pool of volunteers is a nonrandom sample of the targeted US population and has potentially large selection bias.

To help adjust for selection bias, we use the Behavioral Risk Factor Surveillance System (BRFSS) survey (Centers for Disease Control and Prevention, 2022) as the reference survey. The BRFSS is comprised of annual state-level surveys that are combined into a national representative survey with large state-level observations. In addition to the six quota variables, there are ten demographic and health variables collected in the BRFSS that are also predictive of seropositivity but not used in the quota sampling. After removing observations with missing values on any of the sixteen variables, a total of n_s = 367,165 participants were included in the analysis. The CV of the BRFSS sample weights is $\mathrm{CV}(w_s) = 1.92$.

Table 6.1 shows the sample weighted distribution for the 16 variables in the BRFSS and the covid sample. As expected, the distributions of the six quota variables in the two samples are very close. For the ten demographic and health related variables, most of the distributions differ considerably between the two samples. In general, the covid sample participants tend to be more educated, homeowners, employed and healthier. For example, 84% of the covid sample vs. 29% in weighted BRFSS have a college or higher degree. Hence, selection bias exists in the covid sample, and our aim is to reduce the selection bias in the estimation of undiagnosed SARS-CoV-2 seropositivity.

Table 6.2 shows the ALP estimates of the prevalence of undiagnosed seropositivity with the three balancing scores. As noted, the ALP estimate with $b(\mathbf{x}; \hat{\mathbf{B}}_w)$ detected a 4.65% seropositivity rate, close to the rate of 4.67% detected by $b(\mathbf{x}; \hat{\mathbf{B}}_0, \hat{\boldsymbol{\gamma}}_w^T)$. The corresponding two standard errors are also close (0.78 vs. 0.77). In contrast, the unweighted $b(\mathbf{x}; \hat{\mathbf{B}}_0)$ yields a seropositivity rate of 3.95%, close to the unweighted mean of 3.77%, both are subject to selection bias. It is noteworthy that the adaptive balancing score $b(\mathbf{x}; \hat{\mathbf{B}}_0, \hat{\boldsymbol{\gamma}}_w^T)$ produced stable pseudoweights for the covid sample with $\mathrm{CV}(\hat{w}_c) = 2.24$, close to the 2.25 by the unweighted $b(\mathbf{x}; \hat{\mathbf{B}}_0)$, and both are smaller than $\mathrm{CV}(\hat{w}_c) = 2.33$ produced by the weighted $b(\mathbf{x}; \hat{\mathbf{B}}_w)$.

Table 6.1
Covariate distribution (%) in the Covid sample vs. weighted BRFSS.

	Covid Survey	Weighted BRFSS		Covid Survey	Weighted BRFSS		Covid Survey	Weighted BRFSS
Age Group			Urban/Rural			Flu Vaccinated		
18-44	41.6	42.9	Urban	94.7	93.2	Yes	73.8	51.3
45-69	42.6	41.8	Rural	5.3	6.8	No	26.2	48.7
70-95	15.8	15.2	Children present			Cardiovascular		
Sex			Yes	32.5	34.7	Yes	4.1	9.5
Male	47.4	47.8	No	67.5	65.3	No	95.9	90.5
Female	52.6	52.2	Education			Pulmonary		
Race			<=HS	2.6	39.4	Yes	18.8	18.7
White only	77.5	74.8	College	13.8	31.5	No	81.2	81.3
Black only	9.4	12.6	>=College	83.6	29.1	Immune		
Others	13.1	12.5	Homeowner			Yes	23.4	31.1
Ethnicity			Own	75.2	68.8	No	76.6	68.9
Hispanic	15.9	14.1	Rent	20.2	25.6	Diabetes		
Not Hispanic	84.1	85.9	Other	4.7	5.6	Yes	5.5	11.9
Region			Employment			No	94.5	88.1
Northeast	16.7	17.1	Employed	71.2	57.4	Health Insurance		
Midwest	15.8	17.6	Not in Labor Force	23.8	32.2	Yes	97.4	89.0
Mid-Atlantic	20.8	17.3	Unemployed	5.0	10.4	No	2.6	11.0
South/Central	14.2	15.7						
Mountain/Southwest	15.5	15.3						
West/Pacific	17.0	16.9						

Table 6.2
Undiagnosed seropositivity rate among US adults 04/01-08/04/2020.

	CV($\hat{w}_c$)	Estimates (%)	SE* ($\times 10^{-2}$)
Unweighted	0.00	3.77	0.22
$b(\mathbf{x};\hat{\mathbf{B}}_0)$	2.25	3.94	0.52
$b(\mathbf{x};\hat{\mathbf{B}}_w)$	2.33	4.65	0.78
$b(\mathbf{x};\hat{\mathbf{B}}_0,\hat{\boldsymbol{\gamma}}_w^T)$	2.24	4.67	0.77

*: account for the variability due to estimating $\mathbf{B}$, $\mathbf{B}_0$ or $\boldsymbol{\gamma}$.

7. Conclusion and discussion

In this paper, we examine the exchangeability of the outcome conditional on weighted $b(\mathbf{x};\hat{\mathbf{B}}_w)$ and unweighted $b(\mathbf{x};\hat{\mathbf{B}}_0)$ that are used in existing PS-based weighting/matching methods for nonprobability sample inferences. An adaptive balancing score $b(\mathbf{x};\hat{\mathbf{B}}_0,\hat{\boldsymbol{\gamma}}_w^T)$ is proposed to correct for the potential bias in $b(\mathbf{x};\hat{\mathbf{B}}_0)$ in three steps: 1) estimate unweighted $b(\mathbf{x};\hat{\mathbf{B}}_0)$; 2) estimate bias correction factor $b(\mathbf{x};\hat{\gamma}_w)$; and 3) construct $b(\mathbf{x};\hat{\mathbf{B}}_0,\hat{\boldsymbol{\gamma}}_w^T)=b(\mathbf{x};\hat{\mathbf{B}}_0)+b(\mathbf{x};\hat{\boldsymbol{\gamma}}_w)$, which is a monotone function of the estimated participation propensity.

The basic criterion for choosing balancing score is that it should be finer than, if not equal to, the participation propensity in order to balance the distribution of x between the nonprobability sample and the finite population. Both $b(\mathbf{x};\hat{\mathbf{B}}_0,\hat{\boldsymbol{\gamma}}_w^T)$ and $b(\mathbf{x};\hat{\mathbf{B}}_w)$ produce unbiased and comparably efficient estimates

with $b(\mathbf{x};\hat{\mathbf{B}}_0,\hat{\boldsymbol{\gamma}}_w^T)$ more efficient for quota samples when the reference survey is small or has variable sample weights. Survey practitioners should choose as the reference survey a sample that is sufficiently large with stable sample weights. Note that the efficiency gain by increasing the probability sample size n_s is moderate once $n_s > n_c$.

Two limitations are identified: 1) the adaptive balancing score is constructed by assuming the correctness of the logistic regression propensity model in steps 1 and 2 to obtain the unweighted balancing scores and the bias correction factor. Furthermore, 2) the logistic regression in both steps was assumed to have the same functional form. Accordingly, two extensions can be made for future research: 1) Allowing for a different functional form in step 2, where we model the probability for the reference sample selection, from the functional form assumed in step 1. With known selection variables and the selection probability for each reference survey unit, model diagnostics such as a ROC curve can be implemented to assist in the model selection. 2) Constructing various propensity models. A logistic regression model was fitted to estimate the propensity scores at steps 1 and 2 in Section 4. However, misspecification of the logistic regression model might lead to poorly estimated propensity scores that violate the assumption (2.1), and therefore yield biased estimates. Nonparametric approaches such as machine learning methods can offer alternatives, which relax the assumed parametric model specifications regarding variable selection, functional form, and selection of polynomial terms and multiple-way interactions specified in parametric modeling.

In this paper, we discussed how to construct balancing scores that satisfy the CE assumption so that the outcome distribution is exchangeable between the nonprobability sample and the finite population. Note that the balancing score is a function of observed covariates $\mathbf{x}$ that are collected in both the nonprobability sample C and the reference survey S. If important covariates are missing in S or C, then no matter which balancing score is chosen, the FP mean estimates will be unavoidably biased. Important considerations remain such as which variables need to be collected in both C and S, how will the survey questions be harmonized in C and S data collection, and how can measurement or reporting error be minimized in questionnaire design? How these questions are addressed can be critical to satisfying the CE assumption in PS-based adjustment methods for nonprobability sample analysis. In summary, as required by the conditional exchangeability assumption, it is important to have high-quality reference surveys that collect comprehensive sets of variables with minimal measurement and reporting errors, have sufficiently large sample size, and are well designed with informative and stable sample weights.

Acknowledgements

Acknowledgments go to Dr. Barry Graubard for stimulating discussion and to Dr. Lingxiao Wang for simulation support. This research is financially supported by NIH R03 CA252782.

References

Baker, R., Brick, J.M., Bates, N.A., Battaglia, M., Couper, M.P., Dever, J.A., Gile, K.J. 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.

Beaumont, J.-F. (2020). Are probability surveys bound to disappear for the production of official statistics? *Survey Methodology*, 46, 1, 1-28. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2020001/article/00001-eng.pdf.

Brick, J., and Tourangeau, R. (2017). Responsive survey designs for reducing nonresponse bias. *Journal of Official Statistics*, 33, 735-752. DOI: https://doi.org/10.1515/jos-20170034.

Centers for Disease Control and Prevention (2022). Behavioral Risk Factor Surveillance System: Annual survey data. Atlanta, GA: Centers for Disease Control and Prevention, US Department of Health and Human Services. Retrieved from http://www.cdc. gov/brfss/annual_data/annual_data.htm.

Chen, Y., Li, P. and Wu, C. (2020). Doubly robust inference with nonprobability survey samples. *Journal of the American Statistical Association*, 115(532), 2011-2021.

Elliott, M. (2013). Combining data from probability and non-probability samples using pseudo-weights. *Survey Practice*, 2.

Elliott, M., and Valliant, R. (2017). Inference for nonprobability samples. *Statistical Science*, 32, 249-264.

Groves, R., and Peytcheva, E. (2008). The impact of nonresponse rates on nonresponse bias. *Public Opinion Quarterly*, 72(2), 167-189. DOI: https://doi.org/10.1093/poq/nfn011.

Kalish, H., Klumpp-Thomas, C., Hunsberger, S., Baus, H.A., Fay, M.P., Siripong, N., Wang, J., Hicks, J., Mehalko, J., Travers, J., Drew, M., Pauly, K., Spathies, J., Ngo, T., Adusei, K.M., Karkanitsa, M., Croker, J.A., Li, Y., Graubard, B.I., Czajkowski, L., Belliveau, O., Chairez, C., Snead, K.R., Frank, P., Shunmugavel, A., Han, A., Giurgea, L.T., Rosas, L.A., Bean, R., Athota, R., Cervantes-Medina, A., Gouzoulis, M., Heffelfinger, B., Valenti, S., Caldararo, R., Kolberg, M.M., Kelly, A., Simon, R., Shafiq, S., Wall, V., Reed, S., Ford, E.W., Lokwani, R., Denson, J.-P., Messing, S., Michael, S.G., Gillette, W., Kimberly, R.P., Reis, S.E., Hall, M.D., Esposito, D., Memoli, M.J. and Sadtler, K. (2021). Undiagnosed SARS-CoV-2 seropositivity during the first six months of the COVID-19 pandemic in the United States. *Sci Transl Med*, 13(601), eabh3826.

Kennedy, C., Mercer, A., Keeter, S., Hatley, N., McGeeney, K. and Gimenez, A. (2016). *Evaluating Online Nonprobability Surveys.* Washington, DC: Pew Research Center.

Kern, C., Li, Y. and Wang, L. (2021). Boosted kernel weighting – Using statistical learning to improve inference from nonprobability samples. *Journal of Survey Statistics and Methodology*, 9(5), 1088-1113.

Korn, E.L., and Graubard, B.I. (1999). *Analysis of Health Surveys.* New York: John Wiley & Sons, Inc. DOI: https://doi.org/10.1002/9781118032619.

Lee, S., and Valliant, R. (2009). Estimation for volunteer panel web surveys using propensity score adjustment and calibration adjustment. *Sociological Methods & Research*, 37, 319-343.

Li, Y., and Graubard, B. (2012). Pseudo semiparametric maximum likelihood estimation exploiting gene environment independence for population-based case-control studies with complex samples. *Biostatistics*, 13, 711-723.

Li, Y., Graubard, B. and DiGaetano, R. (2011). Weighting methods for population-based case-control studies with complex sampling. *Journal of the Royal Statistical Society, Series C (Applied Statistics)*, 60, 165-185.

Li, Y., Graubard, B., Huang, P. and Gastwirth, J. (2015). Extension of the Peters–Belson method to estimate health disparities among multiple groups using logistic regression with survey data. *Statistics in Medicine*, 34, 595-612.

Li, Y., Irimata, K.E., He, Y. and Parker, J. (2022). Variable inclusion strategies through directed acyclic graphs to adjust health surveys subject to selection bias for producing national estimates. *Journal of Official Statistics*, 38(3), 1-27.

Mercer, A.W., Kreuter, F., Keeter, S. and Stuart, E.A. (2017). Theory and practice in nonprobability surveys: Parallels between causal inference and survey inference. *Public Opinion Quarterly*, 81, 250-271. DOI: https://doi.org/10.1093/poq/nfw060.

Pinsky, P.F., Miller, A., Kramer, B.S., Church, T., Reding, D., Prorok, P., Gelmann, E., Schoen, R.E., Buys, S., Hayes, R.B. and Berg, C.D. (2007). Evidence of a healthy volunteer effect in the prostate, lung, colorectal, and ovarian cancer screening trial. *American Journal of Epidemiology*, 165(8), 874-881.

Rao, J.N.K. (2021). On making valid inferences by integrating data from surveys and other sources. *Sankhyā B*, 83, 242-272.

Rivers, D. (2007). Sampling for web surveys. Paper presented at the *Joint Statistical Meetings - Section on Survey Research Methods*.

Rosenbaum, P., and Rubin, D. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70, 41-55.

Rubin, D. (1978). Bayesian inference for causal effects: the role of randomization. *The Annals of Statistics*, 6, 34-58.

Scott, A., and Wild, C. (2001). The analysis of clustered case-control studies. *Journal of the Royal Statistical Society Series C*, 50, 389-401.

Shah, B.V. (2004). Comment on "Linearization variance estimators for survey data" by A. Demnati and J.N.K. Rao. *Survey Methodology*, 30, 1, 16. Paper available at https://www150.statcan.gc.ca/n1/pub/12-001-x/2004001/article/6991-eng.pdf.

Stuart, E. (2010). Matching methods for causal inference: A review and a look forward. *Statistical Science: A Review Journal of the Institute of Mathematical Statistics*, 25(1).

Valliant, R. (2020). Comparing alternatives for estimation from nonprobability samples. *Journal of Survey Statistics and Methodology*, 8, 231-263.

Valliant, R., and Dever, J. (2011). Estimating propensity adjustments for volunteer web surveys. *Sociological Methods & Research*, 40, 105-137.

Wang, L., Graubard, B.I., Katki, H.A. 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.

Wang, L., Graubard, B.I., Katki, H.A. and Li, Y. (2022). Efficient and robust propensity-score-based methods for population inference using epidemiologic cohorts. *International Statistical Review*, 90, 146-164.

Wang, L., Valliant, R. and Li, Y. (2021). Adjusted logistic propensity weighting methods for population inference using nonprobability volunteer-based epidemiologic cohorts. *Statistics in Medicine*, 40, 5237-5250. DOI: https://doi.org/10.1002/sim.9122.