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

Survey Methodology

Survey Methodology
50-1

SURVEY
METHODOLOGY

Catalogue No. 12-001-XPB

A JOURNAL
PUBLISHED BY
STATISTICS CANADA

JUNE 2005
•
VOLUME 31
•
NUMBER 1

Statistics Canada Statistique Canada

Canada

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

Catalogue No. 12-001-XPB

A journal published by Statistics Canada

June 2024

•

Volume 50

•

Number 1

Statistics Canada Statistique Canada

Canada

SURVEY METHODOLOGY

A Journal Published by Statistics Canada

Survey Methodology is indexed in The ISI Web of knowledge (Web of science), The Survey Statistician, Statistical Theory and Methods Abstracts and SRM Database of Social Research Methodology, Erasmus University and is referenced in the Current Index to Statistics, and Journal Contents in Qualitative Methods. It is also covered by SCOPUS in the Elsevier Bibliographic Databases.

EDITORIAL POLICY

Survey Methodology usually publishes innovative theoretical or applied research papers, and sometimes review papers, that provide new insights on statistical methods relevant to National Statistical Offices and other statistical organizations. Topics of interest are provided on the journal web site (www.statcan.gc.ca/surveymethodology). Authors can submit papers either to the regular section of the Journal or to the short notes section for contributions under 3,000 words, including tables, figures and references. Although the review process may be streamlined for short notes, all papers are peer-reviewed. However, the authors retain full responsibility for the contents of their papers, and opinions expressed are not necessarily those of the Editorial Board or of Statistics Canada.

Submission of Manuscripts

Survey Methodology is published twice a year in electronic format. Authors are invited to submit their articles **through the *Survey Methodology* hub on the ScholarOne Manuscripts website** (https://mc04.manuscriptcentral.com/surveymeth). For formatting instructions, please see the guidelines provided in the journal and on the web site (www.statcan.gc.ca/surveymethodology). To communicate with the Editor, please use the following email: (statcan.smj-rte.statcan@statcan.gc.ca).

Survey Methodology

A journal published by Statistics Canada

Volume 50, Number 1, June 2024

Contents

Special issue for papers presented at the 29th Morris Hansen Lecture

Survey Methodology, June 2024
Vol. 50, No. 1, pp. 1-2
Statistics Canada, Catalogue No. 12-001-X

Preface to the special issue for papers presented at the 29th Morris Hansen Lecture on the use of nonprobability samples

Partha Lahiri[1]

Neyman's seminal paper transformed survey sampling, leading to widespread adoption of probability sampling and associated design-based methods, particularly within national statistical offices. However, perfect implementation of design-based methods relies on a perfect sampling frame of the target finite population, well-designed samples with known non-zero selection probabilities, no nonresponse, no measurement errors, and the use of sampling weights to correct for unequal probabilities. Under these conditions, consistency of traditional design-based estimators and their variance estimators can be assured for large samples, irrespective of the validity of any model that may have been used to construct the estimators. For large sample sizes, the probability sampling approach is indeed attractive to survey practitioners because the same estimation procedure can be used to handle different kinds of outcome variables without the need to model them separately.

Probability sample surveys encounter challenges, including noncoverage, measurement errors, declining participation rates, and high costs. Conversely, nonprobability surveys like voluntary surveys gain traction due to their convenience and cost-effectiveness. In nonprobability sampling, the selection probability mechanism is unknown. Moreover, often selection probabilities are zero for a subset of finite population units. Thus, traditional design-based methods cannot be used to construct estimates or their uncertainty measures, and one needs to rely on models whose assumptions may not always be verifiable. There is now a growing interest in integrating non-probability data with probability surveys, aiming to mitigate these challenges and leverage the strengths of both approaches.

Due to the increasing importance of nonprobability surveys, the Morris Hansen Lecture committee decided to organize the 29th Morris Hansen Lecture on the topic of "Working with Nonprobability Samples: Assessing and Remediating Bias." The Washington Statistical Society inaugurated the Morris Hansen Lecture Series in 1990, supported by a grant from Westat. Subsequently, the National Agricultural Statistics Service (NASS) joined as a co-sponsor of the event, and since then, has consistently hosted the lecture series nearly every year in Washington, D.C.

Given the ongoing Covid pandemic, the 29th Morris Hansen Lecture was conducted as a virtual event on March 1, 2022. The committee extended invitations to Jean-François Beaumont, Courtney Kennedy, and Yan Li, three esteemed experts in the field of nonprobability surveys, to deliver lectures based on their recent research in this area. This special issue features revised versions of three papers presented in the 29th Morris Hansen Lecture event along with discussions and rejoinders.

The first paper authored by Kennedy, Mercer and Lau investigates the measurement issues associated with nonprobability opt-in surveys, frequently utilized to generate estimates for rare domains due to cost

[1] Partha Lahiri, Joint Program in Survey Methodology, University of Maryland, 1218 Lefrak Hall, College Park, Maryland 20742, U.S.A. E-mail: plahiri@umd.edu.

considerations. Through an extensive benchmarking study, the authors identify population subgroups characterized by significant bias in opt-in surveys, attributing a portion of this bias to bogus responses. Their findings underscore the importance of scrutinizing errors arising from bogus responses in nonprobability surveys, emphasizing the need to address not only selection bias but also the issue of erroneous responses.

The second paper authored by Li examines the conditional exchangeability assumption, which serves as a pivotal assumption in propensity score-based adjustment methods. Specifically, Li explores the validity of the exchangeability assumption under various balancing scores and devises an adaptive balancing score aimed at achieving unbiased estimation of finite population means.

The third paper authored by Beaumont, Bosa, Brennan, Charlebois and Chu represents a significant advancement in the field of inverse probability weighting methods for nonprobability samples aimed at mitigating selection bias. Their research encompasses data integration techniques incorporating both parametric and Classification and Regression Trees (CART) methods, with particular emphasis on accounting for the probability sample design. Of note is their substantial focus on variable selection within the context of their proposed methodologies.

I extend my sincere gratitude to Jean-François Beaumont, Editor of *Survey Methodology*, for graciously agreeing to dedicate a special issue of *Survey Methodology* and inviting me to serve as its Guest Editor. I also express my appreciation to the Morris Hansen lecturers – Jean-François Beaumont, Courtney Kennedy, and Yan Li – for accepting my invitation to contribute papers to this special issue based on their presentations.

Furthermore, I am grateful toauthors of all three papers – Jean-François Beaumont, Keven Bosa, Andrew Brennan, Joanne Charlebois, Kenneth Chu, Yan Li, Courtney Kennedy, Andrew Mercer, Arnold Lau – as well asthe discussants of the papers – Vladislav Beresovsky, J. Michael Brick, Michael R. Elliott, Julie Gershunskaya, Jae Kwang Kim, Yonghyun Kwon, Takumi Saegusa, Aditi Sen, and Changbao Wu – for their insightful commentary on the papers. Their contributions have stimulated valuable discussion and enhanced the depth of the research presented.

I believe that the papers, discussions, and rejoinders will serve as an invaluable reference for future research in this dynamic and challenging field.

Partha Lahiri
Guest Editor

Exploring the assumption that commercial online nonprobability survey respondents are answering in good faith

Courtney Kennedy, Andrew Mercer and Arnold Lau[1]

Abstract

Statistical approaches developed for nonprobability samples generally focus on nonrandom selection as the primary reason survey respondents might differ systematically from the target population. Well-established theory states that in these instances, by conditioning on the necessary auxiliary variables, selection can be rendered ignorable and survey estimates will be free of bias. But this logic rests on the assumption that measurement error is nonexistent or small. In this study we test this assumption in two ways. First, we use a large benchmarking study to identify subgroups for which errors in commercial, online nonprobability samples are especially large in ways that are unlikely due to selection effects. Then we present a follow-up study examining one cause of the large errors: bogus responding (i.e., survey answers that are fraudulent, mischievous or otherwise insincere). We find that bogus responding, particularly among respondents identifying as young or Hispanic, is a significant and widespread problem in commercial, online nonprobability samples, at least in the United States. This research highlights the need for statisticians working with commercial nonprobability samples to address bogus responding and issues of representativeness – not just the latter.

Key Words: Nonprobability; Online surveys; Measurement error; Benchmarking.

1. Introduction

Survey statisticians have long understood that raw samples from commercial online nonprobability panels or marketplaces are apt to be unrepresentative (e.g., Rivers, 2007; Dever, Rafferty and Valliant, 2008). The statistician’s task then is combining the respondents’ answers with auxiliary information to adjust away differences between the sample and the target population. Various approaches have been developed for this purpose (see Elliott and Valliant, 2017 for a review). These approaches work, at least in theory, when the answers from nonprobability respondents are genuine and reasonably accurate. But there is growing evidence that sizable shares of nonprobability respondents provide bogus data, including on variables statisticians use for adjustment.

1.1 Focusing on commercial, online nonprobability surveys

Nonprobability samples can take many forms, such as people recruited from a social media platform or a snowball sample of a rare, at-risk population. In American public opinion polling, however, one general form of nonprobability sampling dominates. Over 80% of U.S. public opinion polling is currently conducted using commercial, online nonprobability panels (Kennedy, Hatley, Lau, Mercer, Keeter, Ferno and Asare-Marfo, 2021). Public opinion surveys using commercial online nonprobability samples almost never feature

[1] Courtney Kennedy, Andrew Mercer and Arnold Lau, Pew Research Center, 1615 L St., NW, Washington D.C., Suite 800, 20036, U.S.A. E-mail: CKennedy@PewResearch.org.

a companion probability-based sample for purposes of calibration, a characteristic distinguishing them from studies statistically blending probability and nonprobability samples (e.g., Elliott and Haviland, 2007).

Since “commercial online nonprobability panel” is cumbersome, we will use the shorthand “opt-in panel”. We urge caution in not conflating such commercial data with other, qualitatively different nonprobability sources, such as those in Beaumont (2022) and Li (2022). Commercial opt-in samples have a set of unique issues, which are the focus of this study.

In Section 2 we provide some background on the nonprobability survey benchmarking literature as well as the bogus respondent literature. In Section 3 we present a new benchmarking study comparing average error levels in six different online survey panels. We observe that errors are particularly large for certain subgroups and that those same subgroups are prone to claiming highly unusual characteristics. In Section 4 we present a follow-up study to tease out whether the claims of unusual characteristics are credible or evidence of bogus responding. In Section 5 we discuss some limitations of these data collections. In Section 6 we provide concluding remarks reflecting on the implications of this research.

2. Background

Two lines of methodological research have developed around data quality in opt-in panels. The first is focused on quantifying the average size of errors and comparing them to the levels in probability-based samples. The second line of research is focused on measurement error in individual survey responses caused by fraudulent, mischievous or otherwise insincere respondents in opt-in panels. Our goal in this paper is to connect these literatures using new data and highlight ways that we see them informing each other.

2.1 Benchmarking literature

Many of the studies focused on representativeness have explored the relative accuracy of nonprobability and probability-based survey estimates compared to available population benchmarks. For example, MacInnis, Krosnick, Ho and Cho (2018) found that weighted estimates from a probability-based online sample had significantly smaller root mean square errors than estimates from six different opt-in panels. Overall, the number of studies finding greater accuracy for probability sample estimates (Malhotra and Krosnick, 2007; Chang and Krosnick, 2009; Yeager, Krosnick, Chang, Javitz, Levendusky, Simpser and Wang, 2011; Szolnoki and Hoffmann, 2013; Erens, Burkill, Couper, Conrad, Clifton, Tanton, Phelps, Datta, Mercer, Sonnenberg, Prah, Mitchell, Wellings, Johnson and Copas, 2014; Sturgis, Baker, Callegaro, Fisher, Green, Jennings, Kuha, Lauderdale and Smith, 2016; Dutwin and Buskirk, 2017; Pennay, Neiger, Lavrakas and Borg, 2018) outnumber those finding similar accuracy in nonprobability estimates (Vavreck and Rivers, 2008; Ansolabehere and Schaffner, 2014) by about five to one.

Other studies in this vein have examined the efficacy of different statistical methods for reducing selection bias in estimates from online opt-in samples. One consistent finding in these studies is that even after employing more sophisticated statistical approaches, such as machine learning or doubly-robust methods, there often remain sizable errors in nonprobability survey estimates (Dutwin and Buskirk, 2017; Mercer, Lau and Kennedy, 2018). For example, Dutwin and Buskirk (2017) note that "advanced techniques such as propensity weighting and sample matching did not improve these (nonprobability) measures, and in some cases made matters worse".

One question raised by this literature is why, even with extensive modeling, do opt-in panel estimates often contain large errors? Is it because the models are mis-specified or missing key covariates, or is the opt-in data flawed in ways immune from statistical modeling correction? In addition to leaving open these questions, the benchmarking literature has another limitation. Most benchmarking studies only consider estimates for the full population (e.g., all U.S. adults). They do not explore the possibility that the accuracy of opt-in estimates may vary between major subgroups (e.g., based on age, race, or ethnicity). Our benchmarking study (presented in Section 3) seeks to address this gap by examining subgroup variation and then using those results to better understand a second body of literature on nonprobability surveys.

2.2 Bogus respondent literature

A separate line of research has documented fraudulent, mischievous, or insincere respondents in commercial opt-in panels. The scale of this problem is alarming. The Insights Association (2022) estimates that researchers should anticipate removing 15% to 25% of opt-in completes due to poor data quality. Geraci (2022) put that rate even higher, noting that "just 10 years ago, researchers would need to remove 5%-10% of all interviews from online samples because of poor quality. That proportion is now in the 35%-50% range". This line of research focuses not on whether opt-in respondents are representative of a broader population, but on whether their answers are credible.

It is increasingly clear that bogus respondents are not merely a nuisance (e.g., adding noise to estimates, requiring replacement interviews). Instead, bogus respondents can lead to highly biased estimates and false conclusions. For example, Litman, Rosen, Rosenzweig, Weinberger-Litman, Moss and Robinson (2021) showed that Centers for Disease Control (CDC) reports of high rates of Americans ingesting bleach to protect against COVID-19 were an artifact of bogus respondents in an opt-in sample. Lopez and Hillygus (2018) found that bogus respondents erroneously inflated estimates of public belief in conspiracies by a factor of two. More recently, Westwood, Grimmer, Tyler and Nall (2022) found that bogus respondents erroneously inflated estimates of support for political violence in the United States by over a factor of two.

While data quality concerns with opt-in samples are not a new phenomenon (e.g., Downes-Le Guin, 2005; Baker, Blumberg, Brick, Couper, Courtright, Dennis, Dillman, Frankel, Garland, Groves, Kennedy, Krosnick, Lavrakas, Lee, Link, Piekarski, Rao, Thomas and Zahs, 2010), research about the extent and impact of insincere responding has accelerated in recent years for several reasons. These include concerns about survey bots (Baxter, 2016; Shanahan, 2018; McDowell, 2019; Puleston, 2019; Geraci, 2022), foreign workers concealing their identity to qualify for U.S. surveys (Kennedy, Clifford, Burleigh, Waggoner and Jewell, 2018; Moss, 2018; Ahler, Roush and Sood, 2019), and people taking a survey multiple times on different devices to evade panel security checks (Ahler et al., 2019; Kennedy et al., 2021). While the companies running commercial opt-in panels are aware of and attempt to address many of these challenges, fraudulent respondents continue to make up a sizeable share of cases in online, opt-in samples.

2.3 Connecting the literatures on commercial nonprobability samples

Statistical approaches developed for nonprobability survey data (e.g., Rivers, 2007; DiSogra, Cobb, Chan and Dennis, 2011; Valliant and Dever, 2011; Valliant, 2020) assume that the task at hand is leveraging on the most effective auxiliary variables. In other words, the problem to be solved is addressing the relevant ways in which the opt-in respondents differ from the target population. Well-established theory states that one can remove selection bias by conditioning on the right set of auxiliary variables (Elliott and Valliant, 2017; Mercer, Kreuter, Keeter and Stuart, 2017; Kohler, Kreuter and Stuart, 2019). There is an implicit assumption in this literature that failure to eliminate error in opt-in estimates simply means we haven't found the right auxiliary variables or they are not available. By this logic, progress comes from finding new sources of auxiliary data or developing statistical methods that can better leverage whatever auxiliary data is available. But this logic only holds if measurement error is nonexistent or small, and in general, the conventional wisdom has been that while satisficing – perhaps the most frequently studied source of measurement error – may be somewhat worse in opt-in surveys, it is not a major contributor to error.

The second line of research discussed raises the possibility that bogus respondents actually introduce much larger measurement error than previously thought. When 15% to 50% of the data collected should be discarded for poor quality, then focusing on auxiliary variables risks missing the bigger picture. Sampling quotas and weighting are unlikely to be effective if the variables they use contain large errors introduced on purpose by bad faith respondents.

To be clear, we do not assert that bogus respondents are the only error source for opt-in panels. Instead, it is far more likely that both bogus respondents (i.e., those answering erroneously) and unrepresentative respondents (i.e., those answering genuinely but who, in aggregate, differ from the target population) contribute to error. Our goal in this study is two-fold: (1) to show how the bogus respondent literature helps explain the benchmarking literature, and (2) highlight the need for statisticians working with opt-in samples to address bogus responding *and* issues of representativeness – not just the latter.

3. Benchmarking study

We initially set out to conduct a benchmarking to examine the extent to which the accuracy of online survey estimates differs by subgroups. The results led us to develop a hypothesis about bogus responding. The benchmarking component was not designed to detect bogus responding, and so we designed a follow-up data collection to address that hypothesis directly. We discuss each data collection in turn.

3.1 Sample design for the benchmarking

The benchmarking study featured samples of U.S. adults from each of six online platforms (Table 3.1). Responding sample sizes ranged from 4,912 to 5,147. Three of the samples came from commercial opt-in panels. The opt-in survey vendors, which routinely use quotas, were provided with the quota targets for age × gender, race × Hispanic ethnicity, education. We computed the quota targets from the 2019 American Community Survey.

Table 3.1
Sample sizes and field dates, by source.

Source	n	Field dates
ABS panel 1	5,027	June 14 - 28, 2021
ABS panel 2	5,147	June 14 - 27, 2021
ABS panel 3	4,965	June 29 - July 21, 2021
Opt-in panel 1	4,912	June 15 - 25, 2021
Opt-in panel 2	4,931	June 11 – 27, 2021
Opt-in panel 3	4,955	June 11 – 26, 2021

Note: “ABS” refers to an online panel that is recruited using probability, address-based sampling. “Opt-in panel” refers to a commercial online nonprobability panel or marketplace.

The three other sources are probability-based survey panels that interview online but recruit panelists offline. Most if not all the panelists in these three sources were recruited using address-based sampling (ABS) from the U.S. Postal Service Computerized Delivery Sequence File. Before adopting ABS recruitment, two of these panels recruited offline using random digit dial samples of telephone numbers.

One of the ABS panel samples was used for this methodological study as well as substantive research that is not part of this project. The substantive research required a larger sample size. For that study, the entire panel was invited to participate, out of which 10,606 panelists completed this questionnaire. We did not want this larger sample size confounding the analysis (i.e., by allowing that one sample to be double the size of the other five). To address this, we drew a stratified random sample from the full panel following the panel’s standard procedure for a target sample size of 5,000 completes. Only respondents that were selected as part of this subsample were included in this study. These are the cases that would have been obtained if only that subsample had been invited to participate. The subsampling process was conducted independently from any analysis of the data. As reported below, all three ABS panels performed in a similar manner, giving us confidence that subsampling from the larger panel did not impair this study by rendering its performance meaningfully different.

ABS panels 1, 2 and 3 had study-specific response rates of 61%, 90%, and 71%, respectively. The cumulative response rate to the surveys (accounting for nonresponse to recruitment, to the current survey and for panel attrition) was 5% for ABS panel 1, 3% for ABS panel 2, and 7% for ABS panel 3. Comparable response rates cannot be computed for the opt-in samples. Ipsos was the data collection firm. A common questionnaire was administered in English or Spanish for all six samples.

Each survey panel has their own approach for weighting a national survey of U.S. adults. This project is focused on data quality, and so it was necessary to avoid having different weighting protocols confound comparisons between the samples. We employed a standard weighting approach for all six samples. For the

ABS samples, the weighting protocol began with panel base weights adjusting for differential probabilities of selection. Because design weights do not exist for the opt-in samples, these cases were assigned a starting weight of 1 and treated as if they were simple random samples.

The second step was to calibrate the starting weights for each sample to a common set of population control totals. The population targets were age × sex, education × sex, education × age, race/ethnicity × education, born inside versus outside the U.S. among Hispanics and Asian Americans, years lived in the U.S., Census region × metro/non-metro, volunteered in past year, voter registration status, political party affiliation, frequency of internet use, and religious affiliation. The first six benchmarks came from the American Community Survey. The next three benchmarks came from the Current Population Survey supplements, and the last three came from Pew Research Center National Public Opinion Reference Survey.

3.2 Benchmarking analysis

In total, 25 benchmark measurements were used in this study. They touch on many different topics including smoking, military service, vehicle ownership, health care coverage, income, social program participation, household composition, and more. The benchmarks were derived from high quality federal sources based either on national surveys or administrative data. None of the 25 benchmark variables were part of the weighting protocol. The full list of benchmarks and sources is provided in the appendix.

To estimate survey error relative to benchmarks, we computed the mean absolute value of the difference between the weighted online survey estimate and the benchmark. For a categorical benchmark variable Y with categories $c = 1, \ldots, k,$ let $\overline{Y}_c$ denote the "true" population value and $\overline{y}_c$ denote the survey estimate for the share of the population belonging to category $c.$ For a given survey, the mean absolute error (MAE) for Y is

$$\text{MAE} = \frac{\sum_{c=1}^{k} \left| \overline{y}_c - \overline{Y}_c \right|}{k}. \tag{3.1}$$

In other words, each panel's MAE reflect a two-step process. First, we averaged the difference between the survey and the benchmark across all the answer categories for the question. Second, we computed the average of those 25 averages. We did this separately for each of the six online panels. This way, all the benchmarks have equal influence in the analysis. We see no theoretical justification to do otherwise. Refusal and "Not sure" responses are not included as a benchmarking category. These responses are, however, reflected in the denominator of the online survey estimates because that reflects actual practice. It is unusual for public opinion researchers to re-base estimates just on those providing an answer other than "Refused" or "Not sure".

The average benchmark deviations (Table 3.2) reveal familiar findings as well as new ones. Consistent with prior studies, the estimated average absolute error is systematically lower in the probability-based samples (3.0 percentage points) relative to the nonprobability samples (6.8 percentage points). Within those two groupings, the samples performed roughly the same. The average absolute errors ranged from 2.6 to 3.5 among the probability ABS samples, and it ranged from 5.9 to 7.3 among the nonprobability opt-in samples.

Table 3.2
Average absolute error in online survey estimates for 25 benchmarks.

	All adults	Ages 18-29	Ages 30-64	Ages 65+	High school or less	Some college	College grad	White	Black	Hispanic
ABS Mean	3.0 (0.08)	4.0 (0.31)	3.2 (0.11)	3.1 (0.16)	4.0 (0.16)	3.1 (0.17)	2.5 (0.13)	2.7 (0.09)	4.4 (0.35)	4.2 (0.34)
Opt-in Mean	6.8 (0.09)	11.6 (0.27)	7.4 (0.12)	3.3 (0.13)	7.2 (0.14)	6.4 (0.17)	7.1 (0.18)	5.9 (0.11)	7.6 (0.27)	11.5 (0.29)
ABS 1	2.6 (0.10)	2.9 (0.32)	2.9 (0.15)	2.7 (0.18)	3.2 (0.19)	2.8 (0.19)	2.3 (0.18)	2.4 (0.12)	4.1 (0.47)	3.4 (0.32)
ABS 2	3.5 (0.11)	5.7 (0.45)	3.5 (0.16)	3.6 (0.22)	4.6 (0.23)	3.6 (0.22)	2.8 (0.15)	3.1 (0.12)	4.9 (0.42)	4.7 (0.42)
ABS 3	2.9 (0.16)	3.3 (0.50)	3.3 (0.21)	2.8 (0.27)	4.2 (0.34)	2.9 (0.30)	2.5 (0.17)	2.7 (0.15)	4.2 (0.55)	4.4 (0.63)
Opt-in 1	7.1 (0.16)	11.9 (0.45)	8.1 (0.21)	3.4 (0.17)	7.0 (0.26)	7.3 (0.25)	7.9 (0.31)	6.4 (0.17)	8.2 (0.44)	11.7 (0.48)
Opt-in 2	7.3 (0.15)	12.8 (0.47)	8.1 (0.22)	3.5 (0.20)	7.3 (0.25)	6.6 (0.28)	8.5 (0.33)	6.6 (0.18)	8.6 (0.41)	11.7 (0.47)
Opt-in 3	5.9 (0.15)	10.2 (0.53)	6.1 (0.19)	3.0 (0.19)	7.3 (0.25)	5.3 (0.31)	5.0 (0.25)	4.8 (0.18)	6.1 (0.44)	11.2 (0.48)

Note: Standard errors are shown in parentheses. Estimates for White and Black adults are based on those who do not identify as Hispanic.

More novel, and perhaps under-appreciated in the field, is the substantial variation by subgroup. For the opt-in samples, the average absolute error in estimates for young adults (ages 18-29) is more than three times larger than the error for adults ages 65 and older (11.6 versus 3.3 percentage points). For the ABS samples, by contrast, the average absolute error for estimates based on younger and older adults is far more similar (4.0 versus 3.1 percentage points).

Table 3.2 shows a similar pattern with respect to Hispanic ethnicity. For the opt-in samples, the average absolute error in estimates for Hispanics is nearly double that for (non-Hispanic) White adults (11.5 versus 5.9 percentage points). For the ABS samples, by contrast, the average absolute error for estimates based on Hispanic and White adults is more comparable (4.2 and 2.7 percentage points).

Put differently, after extensively weighting each of the opt-in samples, the estimates for young adults and for Hispanic adults were off by more than 10 percentage points on average. Those are large errors, and it is not clear from the literature why they are concentrated in those two groups. The opt-in estimates for adults ages 65 and over, for example, are relatively accurate, departing from the benchmarks by only about 3 percentage points. Other demographic variables could be examined, of course. We also carried out this analysis by gender and educational attainment, but the variance across those dimensions was much more muted than the differences by age and ethnicity, and so education and gender are not considered further in this study. Instead, the remaining analysis considers why these errors are so concentrated among those self-identifying as young or Hispanic.

3.3 Benchmarks with the largest errors

To investigate why commercial online nonprobability survey errors are concentrated in certain subgroups, we take a closer look at variables where the errors are largest. Table 3.3 presents weighted estimates for the share of U.S. adults receiving four different government benefits: Supplemental Nutritional Assistance Program (SNAP), Social Security, Unemployment Compensation, and Worker's Compensation. The key finding is not simply that the survey estimates contain error, but the errors are all in one direction. Namely, the commercial online nonprobability surveys contain proportionately too many respondents claiming to receive these benefits. The same pattern is observed for the ABS samples, but the magnitude of the errors is dramatically different.

Table 3.3
Estimates for receipt of four different government benefits.

	SNAP	Social security	Unemployment Compensation	Worker's Compensation
Benchmark	*11.1%*	*21.8%*	*9.3%*	*0.4%*
ABS 1	14.0% (0.57)	25.6% (0.55)	12.4% (0.52)	1.3% (0.20)
ABS 2	19.0% (0.71)	27.7% (0.60)	17.2% (0.68)	2.9% (0.35)
ABS 3	18.4% (0.83)	25.6% (0.63)	13.0% (0.72)	1.6% (0.31)
Opt-in 1	29.9% (0.77)	38.7% (0.74)	18.8% (0.68)	10.0% (0.50)
Opt-in 2	30.0% (0.81)	37.5% (0.72)	21.0% (0.71)	11.8% (0.51)
Opt-in 3	21.7% (0.67)	34.7% (0.73)	16.9% (0.65)	7.6% (0.47)

Note: Standard errors are shown in parentheses. SNAP is the Supplemental Nutritional Assistance Program.

For example, receipt of Worker's Compensation is an incredibly rare characteristic among U.S. adults. The population incidence is less than 1%. But according to the commercial online nonprobability samples, the incidence is closer to 10%. Similarly, for receipt of nutritional assistance, the commercial online nonprobability samples estimate the incidence at 22% to 30%, while the true population rate is just 11%.

These results suggest that commercial online nonprobability respondents are prone to saying "Yes" they have certain characteristics. To further distill that phenomenon, Table 3.4 presents the weighted share of respondents in each sample who claimed to receive at least three of the four government benefits measured in the survey.

Again, it is important to note that receiving at least three of these benefits is incredibly rare (0.1% population incidence). The probability-based ABS samples reflect these dynamics, with estimates for the share of adults receiving three or more of these benefits ranging from 0% to 1%. According to the commercial online nonprobability samples, however, the incidence ranges from 6% to 11%.

Table 3.4
Percentage of adults who self-report receiving at least three of four different government benefits.

	All adults	Ages 18-29	Ages 30-64	Ages 65+	HS or less	Some college	College grad	White	Black	Hispanic
Benchmark	*0.1%*	*0.1%*	*0.1%*	*0.2%*	*0.2%*	*0.1%*	*0.0%*	*0.1%*	*0.2%*	*0.1%*
ABS 1	0.8%	1.1%	0.8%	0.4%	1.2%	0.9%	0.3%	0.3%	2.7%	1.1%
	(0.16)	(0.51)	(0.20)	(0.17)	(0.33)	(0.33)	(0.11)	(0.09)	(0.96)	(0.50)
ABS 2	1.2%	2.0%	1.2%	0.7%	2.2%	1.2%	0.3%	0.4%	3.1%	2.7%
	(0.24)	(0.78)	(0.28)	(0.39)	(0.55)	(0.38)	(0.10)	(0.15)	(1.08)	(0.86)
ABS 3	1.4%	2.4%	1.4%	0.4%	2.0%	1.7%	0.3%	1.0%	2.3%	2.9%
	(0.30)	(1.08)	(0.39)	(0.18)	(0.68)	(0.63)	(0.21)	(0.32)	(1.17)	(1.33)
Opt-in 1	7.8%	17.8%	6.9%	0.8%	5.1%	7.5%	11.1%	4.3%	12.4%	17.2%
	(0.44)	(1.44)	(0.58)	(0.35)	(0.69)	(0.77)	(0.96)	(0.46)	(1.60)	(1.47)
Opt-in 2	9.0%	18.0%	8.9%	0.2%	6.0%	7.7%	13.7%	6.9%	10.8%	16.9%
	(0.42)	(1.42)	(0.59)	(0.12)	(0.60)	(0.85)	(0.87)	(0.48)	(1.47)	(1.47)
Opt-in 3	5.9%	14.7%	4.9%	0.7%	6.9%	5.5%	5.1%	3.1%	6.8%	18.6%
	(0.41)	(1.60)	(0.45)	(0.25)	(0.81)	(0.77)	(0.62)	(0.37)	(1.36)	(1.78)

Note: Standard errors are shown in parentheses. The four government benefits are the Supplemental Nutritional Assistance Program (SNAP), Social Security, Unemployment Compensation and Workers' compensation. Estimates for White and Black adults are based on those who do not identify as Hispanic.

Table 3.4 shows these estimates for the subgroups with the largest errors: young adults and Hispanics. For these groups, the error in the commercial online nonprobability estimates is staggering. According to the commercial online nonprobability samples, about 20% of young adults and 20% of Hispanics receive at least three of these benefits. These results beg our central question: What is more likely – that these young and/or Hispanic nonprobability cases actually have this rare characteristic, or that these cases are misrepresenting themselves (i.e., providing bogus answers). As mentioned above, the answer is critical because statistical techniques for commercial online nonprobability data are premised on the first explanation, even though it strains credulity.

3.4 Removing apparently bogus respondents from estimates

A natural question is how the accuracy evaluation changes if the apparently bogus respondents are dropped from analysis. To examine this, we repeated the benchmarking exercise in Section 3.2, but this time we removed all respondents who reported receiving the four government benefits measured (SNAP, Social Security, Unemployment Compensation, and Worker's Compensation). We then re-weighted each sample according to the procedure described above. This yielded some accuracy improvement but was far from a panacea. The results are provided in the appendix. The average benchmark deviation for the nonprobability samples improved by 21% (from 6.8 percentage points on average to 5.4). The nonprobability estimates for young adults and Hispanics improved, but the average errors remained higher than for other demographic groups (8.6 and 8.3 percentage points on average for young adults and Hispanics, respectively). Removing the apparently bogus respondents did not close the accuracy gap between the probability and nonprobability samples, as the average benchmark deviation for the probability samples (2.8 percentage points) remained smaller.

In sum, removing respondents with an egregiously suspicious response pattern helps rather than hurts accuracy, but we do not view this as a robust solution. Those claiming to have received four disparate government benefits are just one subset of all the possible bogus respondents in the nonprobability samples. Prior research (e.g., Kennedy et al., 2021) indicates that other tests for bogus responding would flag a different, if partially overlapping, set of problematic respondents. Relying on a single test or single response pattern is unlikely to completely diagnose bogus responding.

4. Examining bogus responding directly

Results from the benchmarking study suggest that some opt-in subgroups are prone to giving data that is not credible. However, the benchmarking study was not designed to distinguish credible from non-credible responses. A more direct test of bogus responding would be needed to definitively determine if the patterns in the benchmarking analysis stemmed from "unusual but genuine" opt-in respondents or from bogus respondents. Again, we feel this distinction matters because statistical approaches for opt-in data assume opt-in respondents may be unusual but are genuine.

4.1 Sample design for the follow-up

In February of 2022 we conducted a short (14 question) survey of $n = 569$ U.S. adults using a different opt-in panel from the three in the benchmarking study. We refer to this as the "follow-up survey" out of recognition that its purpose was to follow-up on and further probe intriguing findings from the benchmarking study. If the patterns from the benchmarking study replicated in this fourth, separate opt-in panel, that would be strong evidence of a systemic problem in the U.S. opt-in panel space. No probability-based samples were included in this follow up because neither the benchmarking study nor other studies (e.g., Kennedy et al., 2021) find meaningful levels of bogus respondents in probability-based samples.

The goal of the follow-up study was to determine if purportedly Hispanic and/or young opt-in respondents are prone to saying "Yes" no matter what is asked. To assess this, we selected questions for which a "Yes" answer is not credible. One question asked, "Are you licensed to operate a class SSGN submarine?" and offered Yes or No as responses. A class SSGN is a nuclear-powered U.S. naval submarine equipped with cruise missiles, of which there are only four in operation by the U.S. Navy. With approximately 425,000 active duty and reserve personnel, the entire U.S. Navy comprises less than 0.2% of the U.S. adult population, making the share of U.S. adults qualified to operate such a vessel is approximately 0%. A separate question in the follow-up survey was formatted as a battery asking "which of the following did you do in the past week? Check all that apply". The list of activities included two common activities (watched TV, read a book) and four extraordinarily uncommon activities (purchased a private jet, climbed a peak in the Karakoram Mountains, learned to cook halušky, and played jai alai).

The goal of the follow-up study was to determine how many opt-in respondents would select the non-credible answers and whether that behavior was concentrated among respondents identifying as Hispanics and young adults, as it was in the benchmarking component. Making inferences about the U.S. public was not the research goal, and so this analysis is not weighted.

4.2 Results of the follow-up survey

The follow-up experiment bore out the post-hoc hypothesis from the benchmarking: opt-in respondents identifying as young or Hispanic were prone to giving bogus responses. They are prone to reporting affirmatively that they have some characteristic when that is simply not possible in the aggregate numbers observed. The first column of Table 4.1 shows that overall, 5.3% of the follow-up survey respondents claimed to be licensed to operate a class SSGN nuclear submarine. Echoing the benchmarking study, the incidence of this bogus claim was particularly high among Hispanics (23.7%) and those under age 30 (12.1%).

The pattern was the same for claims of doing at least one of the extremely uncommon activities in the past week. The share reporting at least one extremely uncommon activity in the past week (buying a private jet, climbing the Karakoram Mountains, learning to cook halušky, or playing jai alai) was significantly higher among those age 18 to 29 than those age 30 and over ($t = 2.99$, $p < 0.01$). Likewise, the share reporting at least one extremely uncommon activity in the past week was significantly higher among Hispanics than non-Hispanics ($t = 5.11$, $p < 0.01$). These findings make clear what the benchmarking suggested: that commercial opt-in respondents in these subgroups are prone to giving bogus answers.

Table 4.1
Commercial online nonprobability estimates of extremely rare population characteristics.

	Licensed to operate a nuclear submarine	Extremely low incidence behavior
All adults	5.3%	8.4%
	(0.94)	(1.17)
Ages 18-29	12.1%*	17.2%*
	(3.03)	(3.51)
Ages 30+	3.5%	6.2%
	(0.87)	(1.13)
Hispanic	23.7%*	28.0%*
	(4.41)	(4.66)
Non-Hispanic	1.5%	3.6%
	(0.56)	(0.87)

Note: Standard errors are shown in parentheses. Extremely low incidence behavior was defined as having done any of the following activities in the past week: purchasing a private jet, climbing a peak in the Karakoram Mountains, learning to cook halusky, or playing jai alai. Asterisk (*) indicates that the proportion is significantly higher ($p < 0.01$) than for the complementary group based on 2-tailed t-test assuming unequal variances.

4.3 Implications for statisticians working with opt-in data

These results raise an important question. If Hispanic opt-in cases are clearly answering falsely when self-reporting things like operating a nuclear submarine, might they be answering falsely when they self-report being Hispanic? With opt-in data it is generally not possible to validate respondents' ethnicity. That said, the weight of evidence for bogus responding presented here suggests that all answers from respondents making implausible claims should viewed with skepticism. Indeed, the follow-up survey was a mere 14 questions, eliminating excuses along the lines of "maybe some people get tired near the end of long surveys". That is not what the data show. Instead, we see people claiming Hispanic ethnicity also claiming a series of implausible characteristics. Critically, this replicated in four different commercial nonprobability panels. The simplest explanation, which we also consider to be the most credible, is that some opt-in respondents are prone to saying "Yes" no matter what the question is asking, and this holds for adjustment variables as well as survey outcome variables.

The implication for statisticians working with these data is that some of the variables they use to reduce bias (especially variables measured with a Yes/No format) may contain large errors. Moreover, those errors may be concentrated in certain subgroups, rather than distributed randomly within the responding sample. Consequently, statistical techniques for estimation with commercial opt-in data may not work as well as previously thought. Studies ignoring the existence of bogus respondents in these types of samples are at high risk of overstating the performance of various modeling approaches.

It is less clear why bogus responding is also prevalent among young opt-in respondents. Age is not measured with a Yes/No question, and so we would not expect the same positivity bias that appears to be at play with Hispanic ethnicity. In the benchmarking survey, respondents were asked to select their year of birth from a dropdown menu with years ordered from highest to lowest. In the follow-up study, a binned age variable was provided by the sample vendor. Bogus responders may be simply selecting answers toward the top of the list. It is also possible that the choices are strategic, with bogus responders choosing answers that make them more likely to qualify for a survey or potentially to receive higher incentives. While it seems possible that bogus respondents may in fact skew young, there is little reason to believe that demographics and other auxiliary variables are measured any more accurately than substantive ones. Indeed, the distinction between these two types of variables is only meaningful to statisticians.

5. Limitations

This study addresses one class of nonprobability surveys (i.e., commercial online opt-in panels or marketplaces) in one country (the United States). We would not expect the types of errors observed here to be present in nonprobability samples fielded under qualitatively different circumstances. In commercial nonprobability sources, bogus respondents are rewarded for their bad behavior because they often received incentives with monetary value, but there may be no such reward structure in, for example, a bespoke, offline sample of an at-risk population. We also cannot know from our data whether the findings from this study generalize to commercial nonprobability panels in other countries.

Another limitation of this study concerns the benchmarking analysis presented in Section 3. While benchmarking analysis is useful as a means of evaluating the accuracy of survey estimates, it has its limitations. The benchmarks in this study are drawn from government-funded surveys that are conducted at considerable expense and with great attention to survey quality. But they are surveys nevertheless and subject to some of the same problems facing the online surveys. The surveys used as benchmarks have high response rates, on the order to 60% or more. Accordingly, the risk of nonresponse bias is generally thought to be lower for these surveys, though it still exists. Also relevant is the fact that all surveys, no matter the response rate, are subject to measurement error. Questions asked on government-funded surveys are carefully developed and tested, but they are not immune to some of the factors that create problems of reliability and validity in all surveys. The context in which a question is asked (e.g., the questions that come before it) often affects responses to it. Similarly, all survey items may be subject to some degree of response bias, most notably social desirability bias. Especially when an interviewer is present, respondents may sometimes modify their responses to present themselves in a more favorable light (e.g., by overstating their frequency of voting). All of these factors can affect the comparability of seemingly identical measures asked on different surveys. Assessing the quality of data is an inexact process at best. It is therefore important to bear in mind that benchmarking provides measures of estimated bias and is dependent on the particular set of measures included.

6. Discussion

This study is the first to use benchmarking to identify subgroups where bogus responding is concentrated. It is also the first to demonstrate that high rates of bogus data among young adults and self-identified Hispanics appears to be an industry-wide phenomenon, at least in the United States. Four different commercial nonprobability panels or marketplaces all showed the same pattern. By contrast, opt-in estimates for adults ages 65 and older were relatively accurate (mean absolute error of 3.3 percentage points). This suggests that nonprobability approaches, such as hybrid designs, may be differentially effective depending on the subgroup of interest.

This study also raises the question of whether Hispanic adults are more prone to providing bogus answers than other adults. We do not think that is a credible explanation for the results observed. To our minds, realizing that many of these “Hispanic” cases are, in all likelihood, not actually Hispanic is the single most important finding. Its implications for survey statisticians are profound. The implication is that we should not be relying exclusively on techniques like sample matching, propensity models, or hierarchical regressions to fix errors in commercial nonprobability samples. All such approaches assume that respondents are who they say they are; that their demographic information is measured with little to no error. This study demonstrates that with certain types of nonprobability data (namely commercial, online panels), that is not a safe assumption.

Some researchers working with commercial online data are aware of the threat from bogus respondents and take steps to mitigate it. However, more research is needed around the efficacy of current practices because there is some evidence that they are inadequate. Kennedy et al. (2021) found that 84% of bogus respondents passed an attention check (or “trap” question), 87% passed a check for too-fast response time, and 76% of bogus respondents passed both of those popular data quality checks. More sophisticated detection techniques have been proposed (e.g., Jones, House and Gao, 2015), but they do not appear to be widely adopted.

A related concern is that public reporting of results from commercial nonprobability panels rarely discloses whether and how the threat from bogus respondents was addressed. At a minimum, researchers reporting results based on this type of data should disclose what measures were taken to guard against bogus respondents and to what effect. While some organizations may already provide this information, robust disclosure is far from common. Greater awareness and transparency around the existence of bogus respondents in commercial nonprobability samples may help to reduce instances of erroneous findings (e.g., Litman et al., 2021; Westwood et al., 2022) and promote greater caution when interpreting findings from nonprobability samples.

Acknowledgements

Courtney Kennedy is grateful to the Morris Hansen Award Committee for selecting her to present at the 29th Annual Morris Hansen Lecture. All the authors thank Scott Keeter, the editor, and the referees for their constructive suggestions.

Appendix

Table A.1
Benchmarking variables and source.

Variable	Benchmark Source	Question Wording
English proficiency	2019 American Community Survey	Do you speak a language other than English at home? [Ask if speaks a language other than English at home] How well do you speak English? Very well; Well; Not well; Not at all
Citizenship	2019 American Community Survey	Are you a citizen of the United States?
Parent of child in household	2020 National Health Interview Survey	Are you the parent or guardian of any children under age 18? [Ask if parent or guardian of child under age 18] Are any of those children under 18 now living in your household?
Marital status	2021 Current Population Survey March Supplement	Which of these best describes you? Married; Living with a partner; Divorced; Separated; Widowed; Never been married
Number of adults in household	2019 American Community Survey	How many people, including yourself, live in your household? [Ask if more than one person in household] How many, including yourself, are adults, age 18 and older?
Number of children in household	2019 American Community Survey	How many people, including yourself, live in your household? [Ask if more than one person in household] How many, including yourself, are adults, age 18 and older?
Health insurance	2020 National Health Interview Survey	Are you currently covered by any form of health insurance or health plan?
Retirement account	2021 Current Population Survey March Supplement	At any time during 2020 did you have any retirement accounts such as a 401(k), 403(b), IRA, or other account designed specifically for retirement savings?
Received food stamps	2021 Current Population Survey March Supplement	At any time during 2020, did you or anyone in your household receive benefits from SNAP (the Supplemental Nutritional Assistance Program) or the Food Stamp program, or use a SNAP or food stamp benefit card?
Received social security	2021 Current Population Survey March Supplement	During 2020 did you receive any Social Security payments from the U.S. Government?
High blood pressure	2020 National Health Interview Survey	Have you ever been told by a doctor or other health professional that you had hypertension, also called high blood pressure?
Food allergy	2009-2010 National Health and Nutrition Examination Survey	Do you have any food allergies?
Smoking history	2020 National Health Interview Survey	Have you smoked at least 100 cigarettes in your entire life? [Ask if ever smoked 100 cigarettes] Do you now smoke cigarettes... Every day; Some days; Not at all
Vaping history	2020 National Health Interview Survey	Have you ever used an e-cigarette or other electronic vaping product, even just one time, in your entire life? [Ask if ever used e-cigarette] Do you now use e-cigarettes or other electronic vaping products… Every day; Some days; Not at all

Table A.1(continued)
Benchmarking variables and source.

Variable	Benchmark Source	Question Wording
Moved in last year	2021 Current Population Survey March Supplement	Were you living in this house or apartment 1 year ago?
Type of residence	2019 American Community Survey	Which best describes the building where you currently live? (Include all apartments, flats, etc., even if vacant). A mobile home; A one-family house detached from any other house; A one-family house attached to one or more houses; A building with 2 or more apartments; Boat, RV, van, etc.
Home ownership	2019 American Community Survey	Which of the following describes the house, apartment or mobile home where you live? Owned by you or someone in your household with a mortgage or loan (include home equity loans); Owned by you or someone in your household free and clear (without a mortgage or loan); Rented; Occupied without payment of rent
Number of cars	2019 American Community Survey	How many automobiles, vans, and trucks of one-ton capacity or less are kept at home for use by members of your household?
Job status last week	2021 Current Population Survey March Supplement	Last week, did you do any work either for pay or profit? [Ask if did not work last week or refused] Last week, did you have a job either full or part time? Include any job from which you were temporarily absent.
Work affected by Covid-19	2021 Current Population Survey March Supplement	At any time in the last 4 weeks, were you unable to work because your employer closed or lost business due to the Coronavirus?
Had a job last year	2021 Current Population Survey March Supplement	Did you work at a job or business at any time during 2020?
Union membership	2021 Current Population Survey March Supplement	Are you a member of a labor union or of an employee association similar to a union?
Received unemployment compensation	2021 Current Population Survey March Supplement	At any time during 2020, did you receive any State or Federal unemployment compensation?
Received worker's compensation	2021 Current Population Survey March Supplement	During 2020 did you receive any Worker's Compensation payments or other payments as a result of a job-related injury or illness?
Military/veteran status	2019 American Community Survey	Have you ever served on active duty in the U.S. Armed Forces, Reserves, or National Guard?

Table A.2
Average absolute error in online survey estimates for 25 benchmarks, after removing apparently bogus cases.

	All adults	Ages 18-29	Ages 30-64	Ages 65+	High school or less	Some college	College grad	White	Black	Hispanic
ABS Mean	2.8	3.7	3.1	3.0	3.7	2.9	2.5	2.7	4.1	3.8
	(0.08)	(0.30)	(0.10)	(0.17)	(0.16)	(0.18)	(0.13)	(0.09)	(0.37)	(0.32)
Opt-in Mean	5.4	8.6	6.1	3.3	6.2	5.3	5.0	4.9	6.4	8.3
	(0.09)	(0.25)	(0.12)	(0.14)	(0.14)	(0.17)	(0.14)	(0.10)	(0.24)	(0.24)
ABS 1	2.5	2.7	2.8	2.7	3.1	2.7	2.2	2.4	3.7	3.2
	(0.09)	(0.33)	(0.13)	(0.18)	(0.19)	(0.18)	(0.18)	(0.12)	(0.45)	(0.30)
ABS 2	3.3	5.3	3.4	3.5	4.3	3.5	2.8	3.0	4.6	4.3
	(0.11)	(0.42)	(0.16)	(0.22)	(0.24)	(0.23)	(0.16)	(0.13)	(0.38)	(0.40)
ABS 3	2.6	3.2	3.0	2.8	3.7	2.6	2.5	2.6	3.9	3.8
	(0.15)	(0.43)	(0.19)	(0.28)	(0.31)	(0.29)	(0.18)	(0.15)	(0.60)	(0.57)
Opt-in 1	5.8	8.5	6.9	3.3	6.2	6.1	5.5	5.6	6.6	8.2
	(0.15)	(0.39)	(0.21)	(0.18)	(0.26)	(0.23)	(0.25)	(0.17)	(0.38)	(0.40)
Opt-in 2	5.6	9.4	6.4	3.6	6.3	5.3	5.6	5.0	7.4	8.5
	(0.15)	(0.41)	(0.21)	(0.19)	(0.27)	(0.29)	(0.23)	(0.16)	(0.40)	(0.44)
Opt-in 3	4.8	7.8	5.2	3.0	6.2	4.4	3.9	4.2	5.2	8.2
	(0.15)	(0.44)	(0.19)	(0.18)	(0.23)	(0.29)	(0.23)	(0.16)	(0.41)	(0.41)

Note: Apparently bogus cases are defined here as those who claimed to have received all four government benefits measured.

References

Ahler, D.J., Roush, C.E. and Sood, G. (2019). The micro-task market for lemons: Data quality on Amazon's Mechanical Turk. Presented at the Annual Meeting of the Midwest Political Science Association, April 6, 2019.

Ansolabehere, S., and Schaffner, B. (2014). Does survey mode still matter? findings from a 2010 multi-mode comparison. *Political Analysis*, 22(3), 285-303.

Baker, R., Blumberg, S.J., Brick, J.M., Couper, M.P., Courtright, M., Dennis, J.M., Dillman, D., Frankel, M.R., Garland, P., Groves, R.M., Kennedy, C., Krosnick, J., Lavrakas, P.J., Lee, S., Link, M., Piekarski, L., Rao, K., Thomas, R.K. and Zahs, D. (2010). Research synthesis: AAPOR report on online panels. *Public Opinion Quarterly*, 74, 711-81.

Baxter, K. (2016). On the internet, nobody knows you're a bot participant: How bots are contaminating online research data and how we can stop them. Medium. Available at https://medium.com/salesforce-ux/on-the-internet-nobody-knows-youre-a-bot-participant-327dd0da5ce7/.

Beaumont, J.-F., Bosa, K., Brennan, A., Charlebois, J. and Chu, K. (2024). Handling non-probability samples through inverse probability weighting with an application to Statistics Canada's crowdsourcing data. *Survey Methodology*, 50, 1, 77-106. Paper available at http://www.statcan.gc.ca/pub/12-001-x/2024001/article/00004-eng.pdf.

Chang, L., and Krosnick, J.A. (2009). National surveys via RDD telephone interviewing versus the internet: Comparing sample representativeness and response quality. *Public Opinion Quarterly*, 73, 4, 641-678.

Dever, J., Rafferty, A. and Valliant, R. (2008). Internet surveys: Can statistical adjustments eliminate coverage bias? *Survey Research Methods*, 2, 47-62.

DiSogra, C., Cobb, C., Chan, E. and Dennis, J.M. (2011). Calibrating non-probability internet samples with probability samples using early adopter characteristics. *Joint Statistical Meetings, Proceedings of the Survey Methods Section*, 4501-15. Alexandria, VA: American Statistical Association.

Downes-Le Guin, T. (2005). Satisficing behavior in online panels. Presentation at the MRA Annual Conference & Symposium, Chicago, IL, USA. http://www.sigmavalidation.com/tips/05_06_02_Online_Panelists.pdf.

Dutwin, D., and Buskirk, T.D. (2017). Apples to oranges or gala versus golden delicious? Comparing data quality of nonprobability internet samples to low response rate probability samples. *Public Opinion Quarterly*, 81, 213-239.

Elliott, M.N., and Haviland, A. (2007). Use of a web-based convenience sample to supplement a probability sample. *Survey Methodology*, 33, 2, 211-215. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2007002/article/10498-eng.pdf.

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

Erens, B., Burkill, S., Couper, M.P., Conrad, F., Clifton, S., Tanton, C., Phelps, A., Datta, J., Mercer, C.H., Sonnenberg, P., Prah, P., Mitchell, K.R., Wellings, K., Johnson, A.M. and Copas, A.J. (2014). Nonprobability web surveys to measure sexual behaviors and attitudes in the general population: A comparison with a probability sample interview survey. *Journal of Medical Internet Research*, 16(12), e:276.

Geraci, J. (2022). Poll-Arized: Why Americans Don't Trust the Polls and How to Fix Them Before It's Too Late. Houndstooth Press, 153.

Insights Association (2022). Online sample fraud: Causes, costs, and cures. Online, February 11, 2022.

Jones, M.S., House, L.A. and Gao, Z. (2015). Respondent screening and revealed preference axioms: Testing quarantining methods for enhanced data quality in web panel surveys. *Public Opinion Quarterly*, 79, 687-709.

Kennedy, R., Clifford, S., Burleigh, T., Waggoner, P. and Jewell, R. (2018). How Venezuela's economic crisis is undermining social science research − About everything. *Washington Post*, November 7, 2018. Available at https://www.washingtonpost.com/news/monkey-cage/wp/2018/11/07/how-the-venezuelan-economic-crisis-is-undermining-social-science-research-about-everything-not-just-venezuela/.

Kennedy, C., Hatley, N., Lau, A., Mercer, A., Keeter, S., Ferno, J. and Asare-Marfo, D. (2021). Strategies for detecting insincere respondents in online polling. *Public Opinion Quarterly*, 85, 1050-1075.

Kohler, U., Kreuter, F. and Stuart, E.A. (2019). Nonprobability sampling and causal analysis. *Annual Review of Statistics and Its Application*, 6, 149-172.

Li, Y. (2024). Exchangeability assumption in propensity-score based adjustment methods for population mean estimation using non-probability samples. *Survey Methodology*, 50, 1, 37-55. Paper available at http://www.statcan.gc.ca/pub/12-001-x/2024001/article/00008-eng.pdf.

Litman, L., Rosen, Z., Rosenzweig, C., Weinberger-Litman, S.L., Moss, A.J. and Robinson, J. (2021). Did people really drink bleach to prevent COVID-19? A tale of problematic respondents and a guide for measuring rare events in survey data. MedRxiv, DOI: https://doi.org/10.1101/2020.12.11.20246694.

Lopez, J., and Hillygus, D.S. (2018). Why so serious? survey trolls and misinformation. Presented at the Annual Meeting of the Midwest Political Science Association, Chicago.

MacInnis, B., Krosnick, J.A., Ho, A.S. and Cho, M. (2018). The accuracy of measurements with probability and nonprobability survey samples: Replication and extension. *Public Opinion Quarterly*, 82, 707-744.

Malhotra, N., and Krosnick, J. (2007). The effect of survey mode and sampling on inferences about political attitudes and behavior: Comparing the 2000 and 2004 ANES to internet surveys with nonprobability samples. *Political Analysis*, 15(3), 286-323.

McDowell, B. (2019). Minimizing the impact of survey bots. Quirk's Media. Available at https://www.quirks.com/articles/minimizing-the-impact-of-survey-bots.

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.

Mercer, A., Lau, A. and Kennedy, C. (2018). For weighting online opt-in samples, what matters most? Available at https://www.pewresearch.org/methods/2018/01/26/for-weighting-online-opt-in-samples-what-matters-most/.

Moss, A. (2018). After the bot scare: Understanding what's been happening with data collection on MTurk and how to stop it, CloudResearch. Available at https://www.cloudresearch.com/resources/blog/after-the-bot-scare-understanding-whats-been-happening-with-data-collection-on-mturk-and-how-to-stop-it/.

Pennay, D.W., Neiger, D., Lavrakas, P.J. and Borg, K. (2018). The online panels benchmarking study: A total survey error comparison of findings from probability-based surveys and nonprobability online panel surveys in Australia (2nd ed.). The Australian National University. http://csrm.cass.anu.edu.au/research/publications/methods-research-papers.

Puleston, J. (2019). Panel Hacking. Presented at the Annual Conference of the Association for Survey Computing. Available at https://ascconference.org/wp-content/uploads/2019/04/11-Jon-Puleston-Panel-hacking-ASC-2019.pdf.

Rivers, D. (2007). Sampling for web surveys. Paper presented at the 2007 Joint Statistical Meetings, Salt Lake City, UT, USA.

Shanahan, T. (2018). Are you paying bots to take your online survey? Fors Marsh Group. Available at https://www.forsmarshgroup.com/knowledge/news-blog/posts/2018/march/are-you-paying-bots-to-take-your-online-survey/.

Sturgis, P., Baker, N., Callegaro, M., Fisher, S., Green, J., Jennings, W., Kuha, J., Lauderdale, B. and Smith, P. (2016). Report of the inquiry into the 2015 British general election opinion polls. London: Market Research Society and British Polling Council.

Szolnoki, G., and Hoffmann, D. (2013). Online, face-to-face and telephone surveys − Comparing different sampling methods in wine consumer research. *Wine Economics and Policy*, 2, 57-66.

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

Valliant, R., and Dever, J.A. (2011). Estimating propensity adjustments for volunteer web surveys. *Sociological Methods & Research*, 40(1), 105-37.

Vavreck, L., and Rivers, D. (2008). The 2006 Cooperative Congressional Election Study. *Journal of Elections, Public Opinion & Parties*, 18(4), 355-366.

Westwood, S.J., Grimmer, J., Tyler, M. and Nall, C. (2022). Current research overstates American support for political violence. *Proceedings of the National Academy of Sciences*, 119(12).

Yeager, D.S., Krosnick, J.A., Chang, L., Javitz, H.S., Levendusky, M.S., Simpser, A. and Wang, R. (2011). Comparing the accuracy of RDD telephone surveys and internet surveys conducted with probability and non-probability samples. *Public Opinion Quarterly*, 75, 709-747.

Survey Methodology, June 2024
Vol. 50, No. 1, pp. 23-26
Statistics Canada, Catalogue No. 12-001-X

Comments on “Exploring the assumption that commercial online nonprobability survey respondents are answering in good faith”

J. Michael Brick[1]

Abstract

Nonprobability samples are quick and low-cost and have become popular for some types of survey research. Kennedy, Mercer and Lau examine data quality issues associated with opt-in nonprobability samples frequently used in the United States. They show that the estimates from these samples have serious problems that go beyond representativeness. A total survey error perspective is important for evaluating all types of surveys.

Key Words: Total survey error; Fit-for-use; Fabrication.

Kennedy, Mercer and Lau (KML) make a valuable contribution to our understanding of the quality of estimates from opt-in panels. The clarity of the article is noteworthy since much of the literature associated with opt-in panels is related to selection bias and is so technically sophisticated that it may hinder practitioners from fully appreciating the key assumptions and their implications.

KML’s research was intended to examine the accuracy of opt-in estimates for producing domain estimates. The ability to produce estimates for rare domains is a potentially significant benefit of opt-in panels because costs are so much lower for these samples. Most earlier studies had not looked closely at domain estimates. KML’s findings reveal disturbing features of opt-in panels that are consistent with existing literature but may not be well-known to many users of opt-in panels. In essence, large fractions of opt-in interviews are of such poor quality that some researchers suggest dropping 15 to 50 percent of all interviews. I can only imagine what Dr. Deming would have thought of this solution to the problem where the supplier hides the “defects” and it is the customer’s job to inspect and remove them. The typical Deming approach of working on the process or system to avoid the defects in the first place is not feasible for opt-in panels because the process is a black box that is unavailable to the customer.

Their initial study led KML to develop hypotheses about bogus respondents that could be tested in a smaller follow-up study. The follow-up survey clearly shows that a substantial portion of the respondents in opt-in panels are not credible respondents. Here again, they provide the evidence that users of opt-in panels should not ignore.

Before discussing some of the details, a subtext of KML is that the “assumption” that respondents in opt-in panels behave like respondents in probability samples is just an assumption – and a faulty assumption. For decades, surveys conducted with probability samples have been examined for multiple sources of error using the Total Survey Error (TSE) framework. Does it make sense to ignore this framework just because

[1] J. Michael Brick, Statistics and Data Science, Westat, 1600 Research Blvd., Rockville, MD 20850 U.S.A. E-mail: mikebrick@westat.com.

the source of the sample is different? Traditional criteria like coverage and response rates in TSE may not directly apply to opt-in surveys, but many other error sources do.

The most recent American Association of Public Opinion Research report (AAPOR, 2022) on quality metrics reveals the struggle to identify quality online samples is ongoing although progress is limited. Users have little reason to be comfortable with the quality of any opt-in panel. KML only briefly discuss the relative merits of their different opt-in panels. The search for a high-quality opt-in panel is something that many pursued for over a decade, but the evidence thus far suggests this is a chimera.

KML do not attempt a comprehensive review of all sources of error for opt-in surveys, but they clearly show that traditional survey assumptions may not apply to opt-in panels. Speed and low-cost data collection – the main advantages of opt-in panels – should not be the only criteria when deciding on how to conduct a survey. Accuracy of the estimates matters, or at least it should matter! Two important messages in KML are that (1) opt-in panels perform poorly compared to ABS probability-based panels in terms of the accuracy of estimates for domains, and (2) one of the reasons for the poor quality of opt-ins panels is the presence of bogus respondents. Every potential customer of an opt-in panel should understand these facts before deciding whether an opt-in panel is an appropriate source of data.

An opt-in panel could be fit-for-use if all that is needed is a general idea about the size of an estimate (“is it bigger than a breadbox?”). But customers should be very aware that increasing the sample size of the opt-in panel will not improve the accuracy of the estimates because bias does not decrease with sample size. Furthermore, KML show that increasing sample sizes to produce accurate domain estimates is not effective. The presence of bogus respondents adds noise that distorts estimates for domains.

Assumptions

KML show that sampling and selection biases are not the only differences between probability sample surveys and opt-in panel surveys. Let’s begin by considering respondents. Two active lines of research on the quality of respondents in opt-in panels focus on concerns regarding professional respondents, or more generally panel conditioning (are opt-in panel respondents conditioned by responding to other surveys?), and validity (are respondents who they say they are?). Hillygus, Jackson and Young (2014) and Baker, Miller, Kachhi, Lange, Wilding-Brown and Tucker (2014) are examples of this work. However, many users act as if opt-in panel respondents are like probability sample respondents. KML show this assumption is unjustified.

Why wouldn’t respondents to opt-in panels behave like respondents to probability samples? Perhaps the better question is why would we ever believe they are similar? Opt-in panel respondents choose to join a panel or respond without being asked to respond to a specific survey. The motivations of probability sample respondents and opt-in panel respondents are likely to be very different (Keusch, Batinic and Mayerhofer, 2014). Those who argue that choosing to respond to a panel is like choosing to respond to direct survey request in a probability sample are making a large and unsubstantiated assumption. My unverified assumption is that the act of reaching out and actively requesting a household to join is perhaps the major quality advantage that probability-based panels have over opt-in panels. Similarly, the literature of panel conditioning from probability samples may not have much relevance to opt-in panels (or probability-based panels) because that literature examines the effects on responses over time to the same general set of questions rather than every survey request being different.

In Table 3.3 KML provide striking evidence of much greater bias in the opt-in panels than the probability-based panels. Suppose we assume a simple measurement error model where the survey response is subject to error but the benchmark is error-free. This model is

$$y_{s,i} = \mu_i + \varepsilon_{s,i},$$

where $y_{s,i}$ is the 0-1 response to having the government benefit for survey s and respondent i, μ_i is the true value for respondent i, and $\varepsilon_{s,i}$ is the error in survey s for respondent i. In this simple case, the measurement error bias is

$$\text{bias}_{\text{me}} = (1-\mu)\gamma_{\text{FP}} - \mu\gamma_{\text{FN}},$$

where μ is the finite population mean of the μ_i, γ_{FP} is the false positive probability, and γ_{FN} is the false negative probability. Table 3.3 shows the opt-in panel absolute biases for the 4 different government benefits range from 1.3 to 5.1 times higher than those for the ABS samples suggesting poor quality for the opt-in panels.

If we assumed $\gamma_{\text{FP}} = \gamma_{\text{FN}}$, we might think these findings are reasonable, but there is a substantial literature on these rates from probability samples that show γ_{FP} is *negligible* compared to γ_{FN}. As a result, Table 3.3 in KML also raises serious questions about the overestimation of benefits for the probability-based samples. Celhay, Meyer, and Mittag (2022) report the SNAP estimates from the Survey of Income and Program Participation (a longitudinal probability sample) has $\gamma_{\text{FN}} = 0.180$ and $\gamma_{\text{FP}} = 0.013$ resulting in an underestimation of SNAP participation. This finding suggests that the assumption that a probability-based panel behaves like a probability sample may not hold. Survey conditions do matter, and we should be wary of importing an assumption from one setting into another without evidence. The data alone do not allow us to assess whether the unexpected overestimation of benefits in the ABS samples is due to a particular cause or some aggregation of causes. More careful studies of probability-based panels are needed.

Bogus respondents

In the follow-up study, KML confirm their hypothesis that Hispanics and young respondents are much more likely to be bogus respondents. They also rightly question whether these respondents are even Hispanic or young since their responses to these items might also be fabricated.

They suggest that the high rate of fabrication in these subgroups might be related to the motivation to opt into the sample since young and Hispanic respondents are needed for many surveys. Bogus respondents might be motivated to state they are members of these subgroups to obtain incentives or other rewards. This idea seems plausible given the high probability of bogus responses to other “test” items for members of these subgroups.

KML do not discuss bogus respondents in opt-in panels for those who are not members of these subgroups in any depth. But isn’t that the implication from the findings that 3.5 percent of those 35 and older are licensed to operate a nuclear submarine? At the least, the user should be aware of the possibility that a high fraction of *all* responses are bogus.

Is the solution to try to find a way to drop these bogus cases? KML do not endorse that solution and show that even if we wanted to do this, there is no simple way to do it effectively. I completely agree with them. If there is adequate motivation, people will find ways to defeat inspection tools. If opt-in panels wish to be accepted as producing high quality results, then they have considerable work to do. KML have done us a service by clarifying what choices in sample sources really entail.

References

AAPOR (2022). Data Quality Metrics for Online Samples: Considerations for Study Design and Analysis. Downloaded March 13, 2023 https://aapor.org/wp-content/uploads/2023/02/Task-Force-Report-FINAL.pdf.

Baker, R., Miller, C., Kachhi, D., Lange, K., Wilding-Brown, L. and Tucker, J. (2014). Validating respondents’ identity in online samples. *Online Panel Research: Data Quality Perspective, A*, 441-456.

Celhay, P.A., Meyer, B.D. and Mittag, N. (2022). *What Leads to Measurement Errors? Evidence from Reports of Program Participation in Three Surveys*, (No. w29652). National Bureau of Economic Research.

Hillygus, D.S., Jackson, N. and Young, M. (2014). Professional respondents in nonprobability online panels. *Online Panel Research: Data Quality Perspective, A*, 219-237.

Keusch, F., Batinic, B. and Mayerhofer, W. (2014). Motives for joining nonprobability online panels and their association with survey participation behavior. *Online Panel Research: Data Quality Perspective, A*, 171-191.

Comments on “Exploring the assumption that commercial online nonprobability survey respondents are answering in good faith”

Michael R. Elliott[1]

Abstract

Kennedy, Mercer, and Lau explore misreporting by respondents in non-probability samples and discover a new feature, namely that of deliberate misreporting of demographic characteristics. This finding suggests that the “arms race” between researchers and those determined to disrupt the practice of social science is not over and researchers need to account for such respondents if using high-quality probability surveys to help reduce error in non-probability samples.

Key Words: Misreporting; Demographics; Benchmarking.

Kennedy, Mercer, and Lau (KML) are to be commended for an excellent article discussing an underappreciated problem with non-probability samples: the presence of bogus respondents. While theirs is not the only discussion of this topic – Jamieson, Lupia, Amaya, Brady, Bautista, Clinton, Dever, Dutwin, Goroff, Hillygus, Kennedy, Langer, Lapinski, Link, Philpot, Prewitt, Rivers, Vavreck, Wilson and McNutt, 2023 note evidence of similar respondents in studies of ingesting bleach to protect against COVID-19 (Litman, Rosen, Hartman, Rosenzweig, Weinberger-Litman, Moss and Robinson, 2023), belief in conspiracies like PizzaGate (Lopez and Hillygus, 2018), support of political violence (Westwood, Grimmer, Tyler and Nall, 2022), or favorable views of Vladimir Putin (Kennedy, Hatley, Lau, Mercer, Keeter, Ferno and Asare-Marfo, 2021) – they further highlight the intensity of the problem and discover a new facet: intentional misreporting of basic demographics. This latter point is quite important as these quantities, which are often known reasonably accurately for populations of interest from the US Census or other government survey sources, are often used for calibration, to mitigate selection and/or non-response bias.

I and others have advocated for careful and continued funding of a set of major high quality surveys to serve as benchmarks to calibrate non-probability surveys (Wu, 2022). KML suggest such studies might also need to include measures to help detect bogus respondents, in addition to whatever steps might be taken to directly identify such respondents in the non-probability survey itself, such as inclusion of bogus identification questions (Petzel, Johnson and McKillip, 1973; Chandler, Rosenzweig, Moss, Robinson and Litman, 2019). Given that non-probability surveys can easily become the target of individuals trying to influence study outcomes for a wide variety of reasons, we have probably not seen the last of this growing “arms race” between researchers and those determined to disrupt the practice of social science.

[1] Michael R. Elliott, Department of Biostatistics, University of Michigan, M4124 SPH II, 1415 Washington Heights, Ann Arbor, MI 48109. E-mail: mrelliot@umich.edu.

References

Chandler, J., Rosenzweig, C., Moss, A.J., Robinson, J. and Litman, L. (2019). Online panels in social science research: Expanding sampling methods beyond Mechanical Turk. *Behavior Research Methods*, 51, 2022-2038.

Jamieson, K.H., Lupia, A., Amaya, A., Brady, H.E., Bautista, R., Clinton, J.D., Dever, J.A., Dutwin, D., Goroff, D.L., Hillygus, D.S., Kennedy, C., Langer, G., Lapinski, J.S., Link, M., Philpot, T., Prewitt, K., Rivers, D., Vavreck, L., Wilson, D.C. and McNutt, M.K. (2023). Protecting the integrity of survey research. *PNAS Nexus*, 2, 3, pgad049.

Kennedy, C., Hatley, N., Lau, A., Mercer, A., Keeter, S., Ferno, J. and Asare-Marfo, D. (2021). Strategies for detecting insincere respondents in online polling. *Public Opinion Quarterly*, 85, 1050-1075.

Litman, L., Rosen, Z., Hartman, R., Rosenzweig, C., Weinberger-Litman, S.L., Moss, A.J. and Robinson, J. (2023). Did people really drink bleach to prevent COVID-19? A guide for protecting survey data against problematic respondents. *Plos One*, 18, e0287837.

Lopez, J., and Hillygus, D.S. (2018). Why so serious?: Survey trolls and misinformation. *Why So Serious*. Available at SSRN: https://ssrn.com/abstract=3131087.

Petzel, T.P., Johnson, J.E. and McKillip, J. (1973). Response bias in drug surveys. *Journal of Consulting and Clinical Psychology*, 40, 437-439.

Westwood, S.J., Grimmer, J., Tyler, M. and Nall, C. (2022). Current research overstates American support for political violence. *Proceedings of the National Academy of Sciences*, 119, e2116870119.

Wu, C. (2022). Statistical inference with non-probability survey samples. *Survey Methodology*, 48, 2, 283-311. Paper available at https://www150.statcan.gc.ca/n1/pub/12-001-x/2022002/article/00002-eng.pdf.

Vol. 50, No. 1, pp. 29-32
Statistics Canada, Catalogue No. 12-001-X

Comments on "Exploring the assumption that commercial online nonprobability survey respondents are answering in good faith"

Aditi Sen[1]

Abstract

This discussion summarizes the interesting new findings around measurement errors in opt-in surveys by Kennedy, Mercer and Lau (KML). While KML enlighten readers about "bogus responding" and possible patterns in them, this discussion suggests combining these new-found results with other avenues of research in nonprobability sampling, such as improvement of representativeness.

Key Words: Opt-in survey; Measurement error; Data quality; Data integration; Inverse propensity score weighting.

KML in their seminal research focus on the important aspect of measurement error in nonprobability surveys, especially commercial online ones, referred to as "opt-in-surveys". Advanced methods of estimation of population characteristics from nonprobability surveys are commonly developed under the assumption of accuracy of survey responses. In the presence of inaccurate responses, where this assumption is violated, those methods may be inadequate. Thus, when KML question the accuracy of individual survey responses in opt-in surveys, our attention is drawn to this serious issue that calls for further research on the problem.

The most popular type of web survey is based on the so-called "opt-in" or volunteer panel. Unlike probability surveys where a sample, representative of the population, is drawn from a frame, opt-in panels are not constructed using a probability-based design. In such a panel, volunteers are recruited through various convenient but nonprobability methods like quota sampling, snowball sampling etc. and people often join to receive some kind of incentive. Issues like increasing cost and declining response rates of probability surveys are well talked about. In the age of big data and fast and efficient computer programming capabilities, opt-in surveys are receiving much interest. These surveys cost less and panel recruitment as well as receiving responses from volunteers can be achieved quickly. However, there is no guarantee that such samples properly represent the target population. In addition, as KML point out, there is a concern that responses may not be genuine. There is a chance that some respondents might be driven by the incentive or may intentionally provide nonsensical responses. KML highlight these issues of opt-in surveys and find interesting pathways for future research in nonprobability surveys.

Groves and Lyberg (2010) discuss the total survey error framework starting from error typology from Deming's 1944 American Sociological Review article where bias components of error versus variance components of error are clearly noted. It is also noted that earlier sampling theories and methods are most applicable when nonsampling errors are small. Nonsampling errors include nonresponse errors,

[1] Aditi Sen, University of Maryland College Park, USA. E-mail: asen123@umd.edu.

measurement errors and so on that are not related to the sample selection. Hansen, Hurwitz and Bershad (1961) in their paper mention that the collection and processing operations of sample surveys constitute the measurement process and are the sources of measurement error. KML quantify measurement error in individual responses by comparing the average size of errors of opt-in surveys with probability surveys and thus connect two areas of methodological research around data quality in opt-in panels.

KML's paper focuses on the not-so-talked about area of nonsensical responses in opt-in-surveys, which are termed as "bogus responses". The use of benchmarking helps to formulate the hypothesis that these answers to survey questions are nonsensical. This hypothesis is further tested using a "follow-up survey" where questions based on rare events are asked, to which an affirmative response is highly unlikely. In total they work with six surveys; three of them are commercial opt-in surveys where vendors used quota sampling (with the 2019 American Community Survey, ACS, as target) to select the samples. The other three are probability surveys where panelists are recruited using address-based sampling (ABS). In the benchmarking study responses to 25 questions common to all these surveys (treated as estimates) are compared with those from government surveys (treated as true values) like the ACS, the Current Population Survey (CPS) and the National Health Interview Survey (NHIS) in terms of Mean Absolute Error (MAE). A thought along the lines of the follow-up survey is as follows: suppose that the main survey questionnaire could be planned in such a way as to include a set of special "detective" questions (in addition to the main questions). Responses to these questions would be used to measure "probabilities of a bogus response" given covariates. These probabilities of a bogus response could be used to downweight individual responses. For example, we provide the extreme values of the weight adjustment, which is between 0 and 1, using the following conditions: if the probability that a response is bogus is 1 then the weight adjustment is 0, again if the response is well trusted then the weight adjustment is 1. A method could be developed to incorporate these probabilities along with the usual response participation probability weighting, to simultaneously improve representativeness and account for the probability of a bogus answer.

Most interestingly, thankfully to the research of KML, we learn about the presence of patterns in such bogus responding. As the authors indicate, theirs is the first paper that uses benchmarking to identify subgroups that have a high probability of bogus responses. When sub-grouped by demographic variables of respondents, one at a time, primarily age and race-ethnicity, it is observed from the three opt-in-surveys that young (age 18-29 years) and Hispanic respondents are more prone to such trends. Of course, these groupings are subjective; there are other variables like gender, education that are not found to be very impactful in distinguishing those stark differences. It is understandable that considering interaction and subdivision into multiple domains would decrease sample size considerably. In such scenarios one can think of applying small area techniques. Ghosh (2020) provides a great review of different small area models and methods. Here a question can be raised: can a statistical tool be developed, using machine learning methods like regression trees and such, that would help find significant variables to identify bogus responding? This might help to discover the interaction effect between variables unlike considering one at a time, as done by KML. In the context of grouping using machine learning, Loh (2011) reviews some widely available algorithms on classification and regression trees.

Literature on nonprobability surveys has focused on the improvement of representativeness, i.e., to reduce selection bias to make the sample more representative of the population. Much work has been done on the inverse propensity score weighting (IPSW) methods where the propensity is defined by the participation probability of population units in the nonprobability sample. Valliant and Dever (2011), Chen, Li and Wu (2020), Wang, Valliant and Li (2021), Savitsky, Williams, Gershunskaya, Beresovsky and Johnson (2023) derive methods involving combining/stacking nonprobability surveys with probability/reference surveys to estimate participation probabilities, which are otherwise unknown for nonprobability surveys due to their unknown selection mechanism. Here assumptions about ignorability and strict positivity of propensity scores need to be made. These methods generally define a log-likelihood upon creating an indicator variable defining success if a unit is present in the nonprobability sample. The information about the whole finite population being unknown, the subsequent modification of the likelihood into a pseudo-likelihood and use of the reference survey depend on the combining methodology. Researchers thus estimate the propensity score with the help of different data integration techniques and support the performances of estimators in terms of bias and variance with the help of simulation studies and real-life datasets. To be specific, Chen, Li and Wu (2020) compute a doubly robust estimator for finite population means where the name doubly robust comes from two models: one is the propensity score model and the other is the outcome regression model.

To put all these into the context of discussion of the paper by KML, it would be of interest to see the effect of using the ideas developed in the aforementioned papers to estimate the weights for opt-in surveys. Currently the weights in question are developed using calibration to match with population characteristics, but will such estimation procedures involving data integration affect the measurement errors due to bogus responding? KML's paper throws light on the fact that it is not advisable to directly use the responses from opt-in surveys. Researchers should check the credibility of such responses with the help of available probability surveys, like benchmarking with government surveys. For survey statisticians it would be beneficial to know how to properly combine opt-in-surveys with other sources which help validate their credibility and help improve responses or eliminate bogus responding. In essence, the outstanding ideas developed by KML throw the readers into a unique direction of thought, focusing on nonsampling errors. This, combined with the recent development of methodologies on improved representativeness of nonprobability samples, provide us with innovative research outcomes to look forward to.

References

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.

Ghosh, M. (2020). Small area estimation: Its evolution in five decades. *Statistics in Transition*, New Series, Special Issue on Statistical Data Integration, 1-67.

Groves, R.M., and Lyberg, L. (2010). Total survey error: Past, present, and future. *Public Opinion Quarterly*, 74, 5, 849-879. Doi: https://doi.org/10.1093/poq/nfq065.

Hansen, M., Hurwitz, W. and Bershad, M. (1961). Measurement errors in censuses and surveys. *Bulletin of the International Statistical Institute*, 32nd Session, 38, Part 2, 359-74.

Loh, W.-Y. (2011). Classification and regression trees. *WIREs Data Mining Knowl Discov*, 1, 14-23. Doi: https://doi.org/10.1002/widm.8.

Savitsky, T.D., Williams, M.R., Gershunskaya, J., Beresovsky, V. and Johnson, N.G. (2023). Methods for combining probability and nonprobability samples under unknown overlaps. Doi: https://doi.org/10.48550/arXiv.2208.14541.

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

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(4), 5237-5250.

Authors’ response to comments on “Exploring the assumption that commercial online nonprobability survey respondents are answering in good faith”

Courtney Kennedy, Andrew Mercer and Arnold Lau[1]

Abstract

Our comments respond to discussion from Sen, Brick, and Elliott. We weigh the potential upside and downside of Sen’s suggestion of using machine learning to identify bogus respondents through interactions and improbable combinations of variables. We join Brick in reflecting on bogus respondents’ impact on the state of commercial nonprobability surveys. Finally, we consider Elliott’s discussion of solutions to the challenge raised in our study.

Key Words: Commercial nonprobability surveys; Survey panels; Benchmarking studies.

We thank the journal’s leadership for hosting this dialogue and the discussants for offering their thoughtful comments. Each brings a unique perspective. Sen connects our study to other trends in survey statistics. Brick offers sobering reflections on the state of commercial non-probability surveys and helps to situate our work within that. Elliott advances discussion of solutions to the challenge raised in our study.

Sen observes that the demographic groups highlighted in our study were subjective and not exhaustive, as researchers could also look at education, geography, etc. We agree with the overall point and acknowledge that new insights could be gained from casting a wider net for variables correlating with bogus response. We also appreciate her pointing to machine learning as a possible means of identifying bogus respondents through interactions and improbable combinations of variables. The fact that machine learning is scalable and may be adaptive to changing respondent behavior makes it a potentially fruitful avenue for future research. On the other hand, we are skeptical that small area modeling and doubly robust estimators are likely to move the needle on accuracy. Past studies have found that for opt-in samples, such methods offer only marginal improvements over more common calibration methods (Mercer, Lau and Kennedy, 2018; Valliant, 2020). It may be that their limited utility stems from the fact that while such methods are excellent for correcting problems related to selection, they are poorly suited to address the problem of bogus respondents, which is fundamentally about measurement error.

Brick offers several high-level industry reflections that resonate with us. He notes, “The search for a high-quality opt-in panel is something that many pursued for over a decade, but the evidence thus far suggests this is a chimera.” Indeed, our study along with Geraci (2022), Enns and Rothschild (2022) and others suggest that the emergence of such an opt-in panel is growing less likely not more. Previously, statisticians in this space focused on modeling to make commercial nonprobability samples more representative. Now they face the added challenge of determining which interviews are real and which are bogus.

[1] Courtney Kennedy, Andrew Mercer and Arnold Lau, Pew Research Center, 1615 L St., NW, Washington D.C., Suite 800, 20036, U.S.A. E-mail: CKennedy@PewResearch.org.

We also appreciate Brick highlighting the role of the data supplier and how remarkable it is that the client (e.g., the researcher) bears the burden of identifying and remedying the types of errors we document. Our study suggests that data cleaning claims appearing on supplier websites give a false sense of protection from this threat. Indeed, if the suppliers’ quality checks worked, bogus respondent wouldn’t appear in client samples, and studies like ours wouldn’t exist. It is imperative that researchers are aware of the threat posed by bogus respondents, particularly to domain estimates and full population estimates of rare outcomes. In our view, this threat has become so severe that researchers publishing point estimates using commercial non-probability samples should include a fulsome discussion of their approach for dealing with bogus respondents. Journal editors likely have a role in fostering that practice.

One of Brick’s comments specific to our study was particularly intriguing. Reflecting on Table 3.2, he notes how the literature on program participation shows that the likelihood of false negative reporting tends to be significantly higher than the likelihood of false positive reporting. But Table 3.2 shows the opposite pattern in dramatic fashion for opt-in samples and in a more muted but still noticeable fashion for online panels recruited via address-based sampling (ABS). We agree with Brick that this contrarian finding indicates that online panels (both opt-in and ABS-recruited) perform differently than more rigorous probability-based samples on these outcomes. For opt-in panels, we have a reasonably strong hypothesis: bogus respondents tend to report in the affirmative (e.g., “Yes”, “Agree”) regardless of their true status because they want to qualify for future surveys and make more money. For online panels recruited by ABS, however, we are not aware of any hypothesis that would predict false positive reporting. Our suspicion is that the differences between rigorous probability samples and probability-based online panels are not fundamental differences in kind, and are likely a function of mode differences, panel conditioning and other well-known phenomena from the survey methods literature. That being said, we agree with Brick that identifying the precise mechanisms driving these differences seems like fertile ground for theoretical development and future research.

All discussants offered thoughts on possible solutions to the data quality problem explored in our study. As Brick’s remarks suggest, one solution is to simply decide not to use commercial nonprobability samples. While they are undisputedly cheaper and faster, a sizable literature (e.g., Dutwin and Buskirk, 2017; KML; MacInnis, Krosnick, Ho and Cho, 2018; Pennay, Neiger, Lavrakas and Borg, 2018; Yeager, Krosnick, Chang, Javitz, Levendusky, Simpser and Wang, 2011) shows they are less accurate. With Brick, we do not endorse using opt-in samples and assuming one can weed out the bogus cases. We agree with his observation that, given sufficient motivation, bad actors will continue to find ways to circumvent inspection tools in online sources that allow people to opt-in to the process.

Sen raises the possibility of down-weighting respondents found likely to be bogus using detective questions. Prospects for that approach seem to depend on how much of the data provided by the bogus respondents is valid. For nonprobability samples in which the measurement error is likely to stem more from satisficing than fraud, this approach sounds promising. For commercial opt-in samples showing signs of bogus responding (e.g., cases answering “yes” regardless of the question), it is less clear that retaining bogus cases even in a down-weighted capacity would improve mean square errors. Fortunately, these are testable questions, and with Sen, we’d welcome a deeper look into this.

Elliott joins Wu (2022) in advocating for ongoing surveys rigorous enough to produce high quality benchmarks for use in calibrating less rigorous surveys. We enthusiastically second this proposal. At Pew Research Center, we have taken modest steps along these lines, creating an annual, multi-mode address-based survey designed to produce timely benchmark estimates for Americans’ political party affiliation, religious affiliation, and technology use (Pew Research Center, 2022). This multi-modal study reflects the highest rigor we can achieve with our institution’s resources, but much more enhanced designs (e.g., with an in-person stage of data collection) could be possible with the type of investment Elliott proposes. It is clear to us that such new benchmarking studies are needed to improve very low response rate probability-based samples like the three in our study. Whether benchmarking studies can rescue commercial non-probability samples is, to our minds, an open question given the challenge posed by respondents intentionally misreporting their status on both weighting and outcome variables.

References

Dutwin, D., and Buskirk, T.D. (2017). Apples to oranges or gala versus golden delicious? comparing data quality of nonprobability internet samples to low response rate probability samples. *Public Opinion Quarterly*, 81, 213-239.

Enns, P., and Rothschild, J. (2022). Do you know where your survey data come from? Outsourcing data collection poses huge risks for public opinion. Medium, available at https://medium.com/3streams/surveys-3ec95995dde2.

Geraci, J. (2022). *Poll-arized: Why Americans Don’t Trust the Polls and How to Fix Them Before It’s Too Late*. Houndstooth Press, p. 153.

MacInnis, B., Krosnick, J.A., Ho, A.S. and Cho, M. (2018). The accuracy of measurements with probability and nonprobability survey samples: Replication and extension. *Public Opinion Quarterly*, 82, 707-744.

Mercer, A., Lau, A. and Kennedy, C. (2018). For weighting online opt-in samples, what matters most? *Pew Research Center.* http://www.pewresearch.org/2018/01/26/for-weighting-online-opt-in-samples-what-matters-most/.

Pennay, D.W., Neiger, D., Lavrakas, P.J. and Borg, K. (2018). The online panels benchmarking study: A total survey error comparison of findings from probability-based surveys and nonprobability online panel surveys in Australia (2e Eds.) The Australian National University. https://csrm.cass.anu.edu.au/research/publications/methods-research-papers?search_term=The+online+panels+benchmarking.

Pew Research Center (2022). National Public Opinion Reference Survey (NPORS). Available at https://www.pewresearch.org/methods/fact-sheet/national-public-opinion-reference-survey-npors/.

Valliant, R. (2020). Comparing alternatives for estimation from nonprobability samples. *Journal of Survey Statistics and Methodology*, 8(2), 231-263.

Wu, C. (2022). Statistical inference with non-probability survey samples. *Survey Methodology*, 48, 2, 283-311. Paper available at https://www150.statcan.gc.ca/n1/pub/12-001-x/2022002/article/00002-eng.pdf. With discussion available at https://www150.statcan.gc.ca/n1/pub/12-001-x/12-001-x2022002-eng.htm.

Yeager, D.S., Krosnick, J.A., Chang, L., Javitz, H.S., Levendusky, M.S., Simpser, A. and Wang, R. (2011). Comparing the accuracy of RDD telephone surveys and internet surveys conducted with probability and non-probability samples. *Public Opinion Quarterly*, 75, 709-747.

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 $\{y_i, \mathbf{x}_i \colon i \in C\}$ 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 $\overline{Y}_N = \frac{1}{N} \sum_{i \in U} y_i$ 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, $E_C(y \mid U) \neq \overline{Y}_N$, 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 $E(y \mid U) = E_U(\overline{Y}_N)$, 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 $\overline{Y}_N$ 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 $P(i \in C \mid \mathrm{x}, U)$

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., $b(\mathbf{x}_i; \hat{\mathbf{B}}_w) = \hat{\mathbf{B}}_w^T g(\mathbf{x}_i) = \log \hat{p}(i \in C \mid \mathbf{x}_i, U)$. Therefore, under the propensity model (3.1), let $b(\mathbf{x}) = b(\mathbf{x}; \hat{\mathbf{B}}_w)$ in (2.1), that is,

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

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 $\hat{p}(i \in C \mid \mathbf{x}_i, U) = \exp(b(\mathbf{x}_i; \hat{\mathbf{B}}_w))$. 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

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

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

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

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 $\overline{y},$ as if the unit was deleted from the sample. TL variance estimator of $\overline{y}$ is then approximated by (4.4), where $\operatorname{var}\left(\sum_{i \in S} z_i\right)$ accounts for variability due to complex sampling of S. Following Wang et al. (2021), it can be proved that $\overline{y}$ is design-consistent and $\operatorname{var}_{\mathrm{TL}}(\overline{y}) = O\left(\frac{1}{n_c}\right) + O\left(\frac{1}{n_s}\right).$ 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

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

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	$x_1 x_2$	$x_1 x_3$
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. $x_1 x_2$	-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. $x_1 x_2,$ 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. $x_1 x_2,$ 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 $(\overline{Y} = 0.3)$ with the probability and nonprobability sample sizes $n_s = n_c =$ 2,500 and $\mathrm{CV}(w_s) =$ 0.86.

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

Next, we compare in Table 5.3 the two unbiased ALP estimates with $b(\mathbf{x}; \hat{\mathbf{B}}_0, \hat{\gamma}_w^T)$ 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 $\mathrm{CV}(w_s) = 0.38$, 0.86, or 1.50. We make three observations. *First*, as the $\mathrm{CV}(w_s)$ 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 $\mathrm{CV}(w_s)$ increases, the efficiency gain of $b(\mathbf{x}; \hat{\mathbf{B}}_0, \hat{\gamma}_w^T)$ 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 $\mathrm{CV}(w_s)$ 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 $b(\mathbf{x}; \hat{\mathbf{B}}_0, \hat{\gamma}_w^T)$ 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 $\mathrm{CV}(w_s)$, $n_s = n_c = 2{,}500$.

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

Survey Methodology, June 2024
Vol. 50, No. 1, pp. 57-63
Statistics Canada, Catalogue No. 12-001-X

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

Jae Kwang Kim and Yonghyun Kwon[1]

Abstract

Pseudo weight construction for data integration can be understood in the two-phase sampling framework. Using the two-phase sampling framework, we discuss two approaches to the estimation of propensity scores and develop a new way to construct the propensity score function for data integration using the conditional maximum likelihood method. Results from a limited simulation study are also presented.

Key Words: Data integration; Propensity score function; Pseudo weight; Two-phase sampling.

1. Introduction

We would like to congratulate Yan Li for being selected as a Morris Hansen lecturer and for giving an interesting presentation on data integration. Data integration is an emerging area of research to combine multiple data sources in a defensible way. In data integration, by using an independent probability sample as a calibration sample, the selection bias in the convenient sample can be reduced. However, statistical tools for data integration are limited. Thus, I welcome Li’s attempt to develop an additional statistical tool for data integration.

Using the balancing score function to control selection bias in the nonprobability sample is a reasonable idea. How to construct the balancing score function in the context of data integration can be more tricky. Li recognized that the propensity score (PS) estimation method of Chen, Li and Wu (2020) can be inefficient, as the estimation procedure involves using the survey weights in the probability sample. Instead of using weighted estimation, Li proposed an unweighted estimation method and then developed a method for bias correction. The unweighted estimate of PS is also considered by Elliott and Valliant (2017) and has been adopted by some practitioners. In this discussion, we would like to clarify two existing approaches to the estimation of propensity scores and develop a defensible way of constructing the propensity score function for data integration.

The paper is organized as follows. In Section 2, we present a two-phase sampling framework for data integration and the conditional PS model approach is introduced. In Section 3, another approach, called the unconditional model approach, is introduced. The simulation study is presented in Section 4. Some concluding remarks are made in Section 5.

[1] Jae Kwang Kim and Yonghyun Kwon, Department of Statistics, Iowa State University, Ames, Iowa, 50011, U.S.A. E-mail: jkim@iastate.edu.

2. Conditional PS model approach

We use the set-up considered in Yang, Kim and Hwang (2021) where sample A is a probability sample observing $\mathbf{x}$ and sample B is the nonprobability sample observing $(\mathbf{x}, y)$. Table 2.1 presents the general setup of the two sample structures for data integration. As indicated in Table 2.1, sample B is not representative of the target population.

Table 2.1
Data structure for data integration and data fusion.

Data Integration				
Sample	**Type**	X	Y	**Representative?**
A	Probability Sample	✓		Yes
B	Non-probability Sample	✓	✓	No

The formulation is somewhat similar to the two-phase sampling:

1. The first-phase sample $S_1 \equiv A \cup B$ is selected from U and $\mathbf{x}_i$ is observed for all units in S_1.
2. The second-phase sampling $S_2 = B$ is selected from S_1 and y_i is observed for all units in S_2.

Unlike classical two-phase sampling, we do not know the first-order inclusion probability of S_1. Instead, we only know the first-order inclusion probability of the sample A. That is, $\pi_i^{(A)} = P(i \in A \mid i \in U)$ is the (known) first-order inclusion probability of sample A.

Let $\pi_i^{(B)} = P(i \in B \mid i \in U)$ be the (unknown) first-order inclusion probability of sample B. Note that the first-order inclusion probability of S_1 can be written as

$$\begin{aligned} P(i \in S_1 \mid i \in U) &= P(i \in A \cup B \mid i \in U) \\ &= P(i \in A \mid i \in U) + P(i \in B \mid i \in U) - P(i \in A \mid i \in U) P(i \in B \mid i \in U) \\ &= \pi_i^{(A)} + \pi_i^{(B)} - \pi_i^{(A)} \pi_i^{(B)} \end{aligned} \quad (2.1)$$

where the last equality follows from the independence of two samples. Thus, we can express the conditional inclusion probability for the second-phase sample as

$$P(i \in S_2 \mid i \in S_1) = \frac{P(i \in B \mid i \in U)}{P(i \in A \cup B \mid i \in U)} = \frac{\pi_i^{(B)}}{\pi_i^{(A)} + \pi_i^{(B)} - \pi_i^{(A)} \pi_i^{(B)}}. \quad (2.2)$$

Now, since we observe $\mathbf{x}_i$ for $i \in S_1 = A \cup B$, we can make a statistical model for the conditional inclusion probability in (2.2) as a function of $\mathbf{x}$. Let

$$P(i \in S_2 \mid i \in S_1) = p(\mathbf{x}_i; \phi) \quad (2.3)$$

be the statistical model for the conditional inclusion probability with unknown parameter ϕ. We can estimate ϕ by unweighted analysis. That is,

$$\hat{\phi} = \arg\max_{\phi} \sum_{i \in S_1} \left[\delta_i \log p(\mathbf{x}_i; \phi) + \left(1 - \delta_i\right) \log\{1 - p(\mathbf{x}_i; \phi)\} \right],$$

where $\delta_i = \mathbb{I}(i \in B)$ is the indicator function of the event $i \in B$. If a logistic regression model with $\operatorname{logit}\{p(\mathbf{x}_i; \phi)\} = \mathbf{x}_i' \phi$ is used in (2.3), then $\hat{\phi}$ can be obtained by solving

$$\sum_{i \in B} \{1 - p(\mathbf{x}_i; \phi)\} \mathbf{x}_i - \sum_{i \in A} p(\mathbf{x}_i; \phi) \mathbf{x}_i = \mathbf{0}.$$

This unweighted estimation is fully justified, as the conditional inclusion probability model (2.3) is conditional on the first-phase sample $S_1 = A \cup B$. Since the propensity model in (2.3) is conditional on the first-phase sample, it can be called the conditional propensity score (PS) model.

Now, since (2.3) is the model for the conditional inclusion probability in (2.2), we can obtain

$$\frac{\pi_i^{(B)}}{\pi_i^{(A)} + \pi_i^{(B)} - \pi_i^{(A)} \pi_i^{(B)}} = p(\mathbf{x}_i; \hat{\phi}),$$

which implies that

$$\frac{1}{\hat{\pi}_i^{(B)}} = 1 + \frac{1}{\pi_i^{(A)}} \left\{ \frac{1}{p(\mathbf{x}_i; \hat{\phi})} - 1 \right\}. \tag{2.4}$$

Thus, $\hat{w}_i^{(B)} = 1 / \hat{\pi}_i^{(B)}$ in (2.4) can be used as the final pseudo-weight for the elements in sample B.

In practice, we cannot use (2.4) directly as the first-order inclusion probabilities are unknown outside the sample. One way to handle this problem is to estimate $w_i^{(A)} = 1 / \pi_i^{(A)}$ by

$$\tilde{w}_i^{(A)} = E\{w_i^{(A)} \mid \mathbf{x}_i, I_i^{(A)} = 1\} \tag{2.5}$$

following the result of Pfeffermann and Sverchkov (1999). Thus, (2.4) can be changed to

$$\frac{1}{\hat{\pi}_i^{(B)}} = 1 + \tilde{w}_i^{(A)} \left\{ \frac{1}{p(\mathbf{x}_i; \hat{\phi})} - 1 \right\}. \tag{2.6}$$

Li used a parametric model for $E(\pi^{(A)} \mid \mathbf{x}) = \bar{\pi}^{(A)}(\mathbf{x}; \gamma)$ and developed a pseudo maximum likelihood method for estimating γ from the sample. Once $\hat{\gamma}$ is obtained, we can use (2.6) with $\tilde{w}_i^{(A)} = 1 / \tilde{\pi}(\mathbf{x}_i; \hat{\gamma})$.

Instead of using (2.6), Elliott and Valliant (2017) proposed using

$$\frac{1}{\hat{\pi}_i^{(B)}} = \frac{1}{\hat{\pi}_i^{(A)}} \left\{ \frac{1}{p(\mathbf{x}_i; \hat{\phi})} - 1 \right\} \tag{2.7}$$

where

$$\hat{\pi}_i^{(A)} = E\{\pi_i^{(A)} \mid \mathbf{x}_i, I_i^{(A)} = 1\}. \tag{2.8}$$

However, $\tilde{w}_i^{(A)} \neq 1 / \hat{\pi}_i^{(A)}$ in general and the pseudo weight in (2.7) is not theoretically justified.

3. Unconditional PS model approach

Another approach to the PS model is to assume a statistical model for $\pi_i^{(B)} = P(i \in B \mid i \in U)$ such as

$$\pi_i^{(B)} = \pi_B(\mathbf{x}_i; \phi) \tag{3.1}$$

for some parameter ϕ. This unconditional PS model has been considered by Chen et al. (2020) and Wang, Valliant and Li (2021), where the pseudo maximum likelihood method was used to estimate ϕ.

If we wish to improve the efficiency of estimators of ϕ, we can consider the maximum likelihood method as follows. First, if $\pi_i^{(A)}$ are available in S_1, using (3.1), we can derive the following conditional inclusion probability model:

$$\pi_{2i|1}(\phi) = \frac{\pi_B(\mathbf{x}_i; \phi)}{\pi_i^{(A)} + \pi_B(\mathbf{x}_i; \phi) - \pi_i^{(A)} \cdot \pi_B(\mathbf{x}_i; \phi)}. \tag{3.2}$$

In the second step, we can compute the conditional maximum likelihood estimator of ϕ from the combined sample by

$$\hat{\phi} = \arg\max_{\phi} \sum_{i \in S_1} \left[\delta_i \log \pi_{2i|1}(\phi) + (1 - \delta_i) \log\{1 - \pi_{2i|1}(\phi)\} \right], \tag{3.3}$$

where $\pi_{2i|1}(\phi)$ is defined in (3.2). The conditional maximum likelihood estimator in (3.3) is based on the assumption that we can identify the units that belong to the intersection of A and B. Once $\hat{\phi}$ is obtained from the conditional maximum likelihood method, we can use $\hat{w}_i^{(B)} = 1 / \pi^{(B)}(\mathbf{x}_i; \hat{\phi})$ as the pseudo weights for sample B. This conditional maximum likelihood method was also considered by Savitsky, Williams, Gershunskaya, Beresovskyl and Johnson (2022) under the assumption that $\pi_i^{(A)}$ are available in sample B.

If $\pi_i^{(A)}$ are not available outside the sample A, we cannot construct the conditional inclusion probability in (3.2). In this case, we can replace $\pi_i^{(A)}$ by $\tilde{\pi}_i^{(A)} = 1 / \tilde{w}_i^{(A)}$, where $\tilde{w}_i^{(A)}$ is defined in (2.5), and compute

$$\pi_{2i|1}(\phi) = \frac{\pi_B(\mathbf{x}_i; \phi)}{\tilde{\pi}_i^{(A)} + \pi_B(\mathbf{x}_i; \phi) - \tilde{\pi}_i^{(A)} \cdot \pi_B(\mathbf{x}_i; \phi)} \tag{3.4}$$

to apply the above conditional maximum likelihood method in (3.3). The final pseudo weights are given by $\hat{w}_i^{(B)} = 1 / \pi_B(\mathbf{x}_i; \hat{\phi})$ and $\hat{\phi}$ is computed by (3.3).

Instead of the maximum likelihood method, the pseudo weights for sample B can be constructed to satisfy

$$\sum_{i \in B} \frac{1}{\pi_B(\mathbf{x}_i; \phi)} \mathbf{x}_i = \sum_{i \in A} \frac{1}{\pi_i^{(A)}} \mathbf{x}_i. \tag{3.5}$$

Condition (3.5) is often called the calibration property. The calibration property is a desirable property for any pseudo-weights. Once $\hat{\phi}$ is calculated from the calibration equation in (3.5), the final pseudo weight for sample B is given by $\hat{w}_i^{(B)} = 1/\pi_B(\mathbf{x}_i;\hat{\phi})$.

4. Simulation study

A limited simulation study is conducted to compare the performance of estimators, including the methods suggested by the paper of Li. In the simulation, we generate a finite population with $y_i \sim \text{Bernoulli}(p_i)$, $p_i = \text{expit}(-1 + 0.8x_{1i} + 0.2x_{2i} + 0.5x_{1i}x_{2i})$ with (x_1, x_2, x_3) follows from the standared normal distribution. The finite population size is $N =$ 5,000.

From the finite population, sample A is generated repeatedly by the PPS sample with measure of size

$$mos_i = \exp(-1 + 0.5x_{1i} + 0.5x_{3i} - 0.2x_{1i}x_{3i})$$

with sample size n_A = 250. In addition, sample B is selected repeatedly by stratified random sampling with two strata, where stratum 1 is $U_1 = i \in U : x_{1i} > 0$ and stratum 2 is $U_2 = i \in U : x_{1i} \leq 0.$ In stratum 1, $n_{B1} = 0.7n_B$ samples are selected by simple random sampling. In stratum 2, $n_{B2} = 0.3n_B$ samples are selected by simple random sampling. The sample size of B is chosen to be either n_B = 250 or n_B = 2,500 so that the sampling ratio is either 5% or 50%. The design weights for sample A are available in sample A, but not in sample B. The study variable y is available only in sample B. The covariate of the main effects and their pairwise interaction effects $(x_1, x_2, x_3, x_1x_2, x_1x_3, x_2x_3)$ are available in $A \cup B$.

We compare the following estimators:

Mean C Sample mean of the nonprobability sample C. *Unweighted* in the paper.

WBS ALP(Adjusted Logistic Propensity) estimator using weighted balancing score method, proposed by Wang et al. (2021).

ABS ALP estimator using adaptive balancing score method, proposed by Li.

CLW Chen et al. (2020)'s IPW(Inverse Probability Weighting) estimator using logistic regression model for $\pi_i^{(B)}$.

Cal Calibration estimator that satisfies (3.5) using logistic regression model for $\pi_i^{(B)}$.

CPS The proposed pseudo weight estimator (2.6) using the conditional inclusion probability model and the smoothed weights in (2.5). The logistic regression model is used for the conditional inclusion probability model, and Poisson regression was used for smoothing weights of sample A in (2.5).

UCPS The pseudo weight estimator proposed in Section 3 using the logistic regression model for $\pi_i^{(B)}$ with $\hat{\phi}$ estimated by the conditional maximum likelihood method in (3.3).

While the sample B is selected using stratified sampling, the propensity scores of **WBS**, **ABS**, **CLW**, **CPS**, and **UCPS** were fitted from the logistic model, and we allowed model misspecification on the response model of $\pi^{(B)}$.

The simulation results after 1,000 simulation runs are summarized in Table 4.1. When $n_B = 250$, the ABS, the CPS, and the UCPS estimators tend to outperform all other estimators considered. When $n_B =$ 2,500, the CPS and UCPS estimators are better than the other estimators considered. The ABS and WBS methods are developed based on the assumption that the overlap between the two samples is negligible, but this assumption does not hold for $n_B =$ 2,500, as the sampling rate for sample B, $n_B / N = 0.5$, is non-negligible.

Table 4.1
Bias, standard error, and root mean square error after 1,000 repetitions.

	$n_B = 250$			$n_B = 2{,}500$		
	BIAS	**SE**	**RMSE**	**BIAS**	**SE**	**RMSE**
Mean C	0.0533	0.0252	0.0589	0.0514	0.0052	0.0517
WBS	0.0087	0.0275	0.0289	0.0053	0.0139	0.0149
ABS	0.0097	0.0264	0.0281	0.0097	0.0130	0.0162
CLW	0.0084	0.0278	0.0291	-0.0081	0.0234	0.0248
Cal	0.0061	0.0284	0.0291	0.0080	0.0140	0.0161
CPS	0.0095	0.0263	0.0279	0.0035	0.0116	0.0121
UCPS	0.0094	0.0263	0.0280	0.0035	0.0116	0.0121

5. Concluding remark

In constructing pseudo-weights, model assumptions for the nonprobability sample are used. The model assumptions can be classified into two groups, one is the conditional PS model approach and the other is the unconditional PS model approach. The conditional PS model approach is computationally attractive but the smoothing weights for sample A should be constructed correctly. In the unconditional PS model approach, the pseudo maximum likelihood method of Chen et al. (2020) has been used. Li's method is more efficient than the pseudo maximum likelihood method as long as the sampling rate for sample B is negligible. In this paper, we propose an alternative approach using the conditional maximum likelihood method as an efficient estimation method, which can be justified even when the sampling rate for sample B is non-negligible. The computation for the conditional maximum likelihood method is somewhat involved. Beaumont, Bosa, Brennan, Charlebois and Chu (2024) independently proposed a very similar method, which was called the maximum sample likelihood method. Further investigation of the proposed method will be presented elsewhere.

Acknowledgements

We thank Professor Partha Lahiri for the invitation for discussion and two anonymous referees and the Editor for constructive comments. The research was partially supported by a grant from the US National Science Foundation (Grant Number: 2242820) and a grant from the Iowa Agriculture and Home Economics Experiment Station, Ames, Iowa.

References

Beaumont, J.-F., Bosa, K., Brennan, A., Charlebois, J. and Chu, K. (2024). Author's response to comments on "Handling non-probability samples through inverse probability weighting with an application to Statistics Canada's crowdsourcing data": Some new developments on likelihood approaches to estimation of participation probabilities for non-probability samples. *Survey Methodology*, 50, 1, 123-141. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2024001/article/00001-eng.pdf.

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

Elliott, M.R., and Valliant, R. (2017). Inference for nonprobability samples. *Statistical Science*, 32(2), 249-264.

Pfeffermann, D., and Sverchkov, M. (1999). Parametric and semiparametric estimation of regression models fitted to survey data. *Sankhyā B*, 61, 166-186.

Savitsky, T.D., Williams, M.R., Gershunskaya, J., Beresovskyl, V. and Johnson, N.G. (2022). Methods for combining probability and nonprobability samples under unknown overlaps. arXiv:2208.14541.

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(24), 5237-5250.

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, 1, 29-58. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2021001/article/00004-eng.pdf.

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

Causal inference, non-probability sample, and finite population

Takumi Saegusa[1]

Abstract

In some of non-probability sample literature, the conditional exchangeability assumption is considered to be necessary for valid statistical inference. This assumption is rooted in causal inference though its potential outcome framework differs greatly from that of non-probability samples. We describe similarities and differences of two frameworks and discuss issues to consider when adopting the conditional exchangeability assumption in non-probability sample setups. We also discuss the role of finite population inference in different approaches of propensity scores and outcome regression modeling to non-probability samples.

Key Words: Causal inference; Finite population; Non-probability sample.

1. Introduction

I congratulate Professor Yan Li on another important addition to her active research on non-probability samples. In her paper, Professor Li classified existing research on non-probability samples into (1) the propensity score weighting methods and (2) the propensity score matching methods, and discussed that the conditional exchangeability (CE) assumption is required for the former. After reviewing existing methods in view of the CE assumption, Professor Li proposed the novel adaptive balancing score to ensure that the CE assumption holds. As the crystallization of accumulating literature on non-probability samples and causal inference, her paper demands a fair amount of background knowledge in order to understand complex concepts. The focus of our discussion here is to examine basic concepts and foundational issues which Professor Li's sophisticated presentation touched only lightly.

This discussion is organized as follows. In Section 2, we review the conditional exchangeability assumption in causal inference. We describe differences of probabilistic frameworks in causal inference and non-probability samples, and discuss issues to consider when adopting the conditional exchangeability assumption in non-probability samples. In Section 3, we describe two major approaches in missing data problems including causal inference. Then we discuss issues of the role of finite population inference arising from the conditional exchangeability assumption in different approaches.

[1] Takumi Saegusa, Department of Mathematics, University of Maryland, College Park, Maryland 20742, United States of America. E-mail: tsaegusa@umd.edu.

2. Causal inference

First, we discuss the relationship between the CE assumption and causal inference. In the paper, the CE assumption is formulated as the equality

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

where $b(x)$ is a function of covariates x referred as a balancing score, U is a finite population, and $C \subset U$ is a non-probability sample. Though simply defined, the criterion of its choice in the paper indicates that the balancing score seems to be implicitly defined to satisfy the CE assumption. Moreover, it is stated as a fact without much discussion that any quantity (including the propensity score) finer than the propensity score satisfies the CE assumption as a balancing score. An important literature that helps understand these concepts is the one coauthored by Professor Li (Wang, Graubard, Katki and Li, 2022), which is, to the best of our knowledge, the first paper that explicitly introduced balancing scores and conditional exchangeability in causal inference to the non-probability sample literature. In Wang, Graubard, Katki and Li (2022), however, these concepts were directly borrowed from the work of Rosenbaum and Rubin (1983) on causal inference, and results on propensity scores were claimed to hold in the non-probability setting without formal discussion. Because the definitions of the CE assumption and balancing score in the paper are different from those in Rosenbaum and Rubin (1983), and because the counterfactual framework of Rosenbaum and Rubin (1983) is fairly different from the setting of non-probability samples, it is worthwhile to pay a close attention to similarities and differences between causal inference and non-probability samples.

To this end, we first briefly summarize Rosenbaum and Rubin (1983) where variables of interest are potential outcomes $(Y(0), Y(1))$, covariates X, and treatment assignment $Z \in \{0,1\}$. The balancing score $b(x)$ in Rosenbaum and Rubin (1983) was defined as the function of covariates $X = x$ that satisfies the conditional independence between X and treatment assignment Z given $b(X)$ (i.e., $X \perp Z \mid b(X)$). It was shown that the propensity score into treatment is a balancing score, and that any function of x that can get mapped into the propensity score is also a balancing score. As the definition suggests, there is no requirement on the relationship between potential outcomes and covariates. The assumption that connects these variables is the conditional exchangeability with respect to covariates (or strong ignorability of Rosenbaum and Rubin (1983)), defined differently as the conditional independence between the potential outcomes and treatment assignment given covariates (i.e., $(Y(0), Y(1)) \perp Z \mid X$). The main result is that conditional exchangeability with respect to covariate X implies conditional exchangeability with respect to a balancing score $b(X)$. In other words, starting from the key conditional exchangeability assumption given covariates x one can reduce the information of x to a balancing score. Balancing scores $b(x)$ are only meaningful in the presence of conditional exchangeability with respect to covariates x. An implication of this result is that the difference between two potential outcomes are explained only by treatment assignment.

A natural way to apply these results to the non-probability sample setting is to consider selection to the non-probability sample as treatment assignment, and outcomes in the non-probability sample C and the rest in the finite population (i.e., $U \setminus C$) as two potential outcomes. In this setting, the conditional exchangeability of Rosenbaum and Rubin (1983) implies the conditional exchangeability with respect to the propensity score so that C and $U \setminus C$ are comparable given the propensity score. In contrast, Professor Li immediately assumes comparability of C and U given the propensity score. From the causal inference perspective, comparability of Rosenbaum and Rubin (1983) is a consequence of a conceptually checkable assumption while Professor Li begins with the desired comparability by assuming it. If, instead, one starts from conditional exchangeability as in Rosenbaum and Rubin (1983), a result still may not be satisfactory because two samples (i.e., C and $U \setminus C$) remain different by "treatment" of participation in a non-probability sample. For example, if non-probability samples are hospital records or participants of a certain educational program, both samples differ due to receipt of care by the hospital or the educational effect. Even if we do not find such "treatment" that differentiates the non-probability sample and the rest, the conditional comparability between C and $U \setminus C$ does not necessarily correspond to the finite population $U.$ To achieve the correct target population, one needs to obtain a distribution of the propensity score in the finite population $U.$ This task is not simple to carry out as described below in relation to the odds representation of the propensity score.

Another approach is to deviate from causal inference by starting from the conditional independence between Y and selection Z into C given X instead of the conditional exchangeability with potential outcomes. In this case, all derivations in fact remain valid to conclude the result that $Y \perp Z \mid X$ implies $Y \perp Z \mid b(X)$ as desired. However, a new conditional independence assumption is simply the standard missing at random (MAR) assumption in the missing data problem, which is also adopted by Chen, Li and Wu (2020) on their non-probability sample research. The MAR assumption is familiar to many statisticians and easier to examine than the conditional exchangeability assumption of Professor Li. If this approach is the one implicitly adopted in Wang, Graubard, Katki and Li (2022), as well as the current paper, it is worthwhile to discuss additional benefits of this approach over the MAR assumption in addition to the discrepancy between $U \setminus C$ and U for comparability. If a different approach is adopted, an unverified relationship between balancing scores and the CE assumption (2.1) should be explicitly derived. As an aside, we would like to point out that Chen, Li and Wu (2020), is not the only literature that does not use the CE assumption of Professor Li for the propensity score weighting methods (see e.g. Kim and Morikawa (2023) for the non-ignorable missing case).

As mentioned above, the comparability of C and $U \setminus C$ allows reliable estimation of the regression model based on C for items in $U \setminus C$ but the estimation of $\overline{Y}_N$ requires consistent estimation of propensity scores for U to bridge regression given X to the entire population $U.$ However, simple estimation of the propensity score is not possible because X is not available for all items in $U \setminus C.$ The variable X is available in a reference sample S from U with a known design but S is not a simple alternative to $U \setminus C$ because items in S can be also in a non-probability sample $C.$ To address this challenging issue, Wang, Valliant and Li (2021) found the relationship between the propensity score into C relative to U and the propensity score into C relative to the stacked sample of C and U where the same items in C and S are treated differently (for a rigorous derivation, see Savitsky, Williams, Gershunskaya, Beresovsky and Johnson (2023)). Using this relationship, Professor Li modeled the latter propensity score by binary regression to estimate the former. The event for the latter propensity score for a stacked sample is artificially constructed and conceptually difficult to model. This issue enhances the higher possibility of model misspecification, which would invalidate design-consistent estimation of $\overline{Y}_N$. The event for the former propensity score is the original event, and is natural to model. This approach was adopted by Savitsky, Williams, Gershunskaya, Beresovsky and Johnson (2023).

3. Finite population inference

Another concept we want to discuss is the role of the finite population in non-probability samples. The goal of the paper is to develop a design-consistent estimator of the finite population average $\overline{Y}_N$. For design consistency, one assumes series of conditions on the sequence of finite populations with all variables except selection into samples treated non-random. In contrast, the model-based approach treats the finite population as a random realization from the super population, and models the stochastic relationship among variables. In the missing data research, on the other hand, two major approaches (and their combinations) for estimation are the propensity score modeling and the outcome regression modeling. A more suitable approach to the design-based approach is the propensity score modeling that models selection into samples given covariates. This is because one can consider random selections while all other variables can be treated fixed. On the other hand, the outcome regression modeling assumes a distribution for Y given X, and is suitable for the model-based approach.

Professor Li made a difficult attempt to bridge the outcome regression approach to the design-based approach. Note that the conditional expectation can be considered as regression with conditioning variables as covariates. From this view, the approach in the paper seems to be purely the model-based approach based on the outcome regression. However, Professor Li attempted to carefully develop the conditional expectation step by step beginning a finite population and a non-probability sample. If the condition was purely model-based, the variable y in the condition (2.1) is simply a random variable from the super population. In the conditional approach of the paper, this variable y should be clearly defined in relation to the finite population U and the non-probability sample C through indices. If y is a random choice of a variable from a sample S from U, $E_S[y \mid U] = \sum_{i \in U} \pi_i Y_i$ where π_i is the inclusion probability for the unit i. In this case, the self-weighting sample S satisfies $E_S[y \mid U] = \overline{Y}_N$ but a stratified sample S, for example, does not satisfy this equality in general. In other words, the claimed issue of bias may not be unique to a non-probability sample. To fully appreciate the conditional exchangeability condition, a clear definition of y in C and/or U is much desired. Moreover, it is desirable to elucidate how the model-based condition of the CE assumption leads to the design-based result despite conceptual discrepancy.

References

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. Retrieved from https://doi-org.proxy-um.researchport.umd.edu/10.1080/01621459.2019.1677241. Doi: 10.1080/01621459.2019.1677241.

Kim, J., and Morikawa, K. (2023). An empirical likelihood approach to reduce selectionbias in voluntary samples. To appear in *Calcutta Statistical Association Bulletin*, 35.

Rosenbaum, P.R., and Rubin, D.B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1), 41-55. Retrieved from https://doi-org.proxy-um.researchport.umd.edu/10.1093/biomet/70.1.41. Doi: 10.1093/biomet/70.1.41.

Savitsky, T.D., Williams, M.R., Gershunskaya, J., Beresovsky, V. and Johnson, N.G. (2023). *Methods for Combining Probability and Nonprobability Samples Under Unknown Overlaps*.

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(24), 5237-5250. Retrieved from https://doi-org.proxy-um.researchport.umd.edu/10.1002/sim.9122. Doi: 10.1002/sim.9122.

Survey Methodology, June 2024
Vol. 50, No. 1, pp. 71-75
Statistics Canada, Catalogue No. 12-001-X

Author's response to comments on "Exchangeability assumption in propensity-score based adjustment methods for population mean estimation using non-probability samples"

Yan Li[1]

Abstract

In this rejoinder, I address the comments from the discussants, Dr. Takumi Saegusa, Dr. Jae-Kwang Kim and Ms. Yonghyun Kwon. Dr. Saegusa's comments about the differences between the conditional exchangeability (CE) assumption for causal inferences versus the CE assumption for finite population inferences using nonprobability samples, and the distinction between design-based versus model-based approaches for finite population inference using nonprobability samples, are elaborated and clarified in the context of my paper. Subsequently, I respond to Dr. Kim and Ms. Kwon's comprehensive framework for categorizing existing approaches for estimating propensity scores (PS) into conditional and unconditional approaches. I expand their simulation studies to vary the sampling weights, allow for misspecified PS models, and include an additional estimator, i.e., scaled adjusted logistic propensity estimator (Wang, Valliant and Li (2021), denoted by sWBS). In my simulations, it is observed that the sWBS estimator consistently outperforms or is comparable to the other estimators under the misspecified PS model. The sWBS, as well as WBS or ABS described in my paper, do not assume that the overlapped units in both the nonprobability and probability reference samples are negligible, nor do they require the identification of overlap units as needed by the estimators proposed by Dr. Kim and Ms. Kwon.

Key Words: Conditional Exchangeability; Causal inferences; Propensity score; Randomized trials; Observational studies; SARS-CoV-2 seroprevalence study.

I want to thank the discussants for their insightful comments of my paper and for the excellent additional references they cite. I will begin by addressing Dr. Saegusa's discussion on two major points. The first contrasts the differences between the conditional exchangeability (CE) assumption for causal inferences versus the CE assumption for finite population inferences using nonprobability samples. The second point focuses on distinguishing between design-based versus model-based approaches for finite population inference using nonprobability samples.

1. Response to comments by Dr. Saegusa

The CE assumption in causal inference and in finite population inference

Dr. Saegusa provided a thorough explanation of the CE assumption for estimating causal effects of treatments in randomized trials and observational studies. The key condition is that conditional exchangeability is satisfied in order to make causal inferences. In randomized trials, CE of potential outcomes is achieved through random assignment of treatments. Conversely, in the analysis of observational

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

studies, CE is assumed rather than guaranteed to draw causal conclusions. The CE assumption in observational studies asserts that the distribution of potential outcomes (for different treatments), given all *observed* covariates, are exchangeable across treatment groups. In this case, there are no *unobserved* covariates that influence both the treatment assignment and the outcome of interest; see Rubin (2007) for causal effect estimation using randomized trials and observational studies. Note that unlike for *randomized trials*, in the context of finite population (FP) inference using nonprobability samples, CE is presumed rather than guaranteed, which is similar to the CE assumption needed for causal inference in *observational* studies.

Dr. Saegusa adeptly linked self-selection into nonprobability samples to treatment assignment in causal inference, and defined that being self-selected into the non-probability sample C versus in the rest in the finite population (i.e., U\C) as the two treatments. However, in FP inference using nonprobability samples, we are interested in estimating the FP mean for a single outcome, rather than the treatment effects, i.e., the difference between two potential outcomes under the treatments of C and U\C. There are no multiple treatments applied to “like” groups to obtain multiple potential outcomes. Instead, only one single potential outcome is realized. As a result, the CE assumption in FP inference differs from CE in observational studies, where it asserts that the distribution of the (*single*) outcome is exchangeable between the nonprobability sample C and the finite population U, given all *observed* covariates. Under this CE assumption, the FP mean can then be inferred using C without observing the outcome in U\C.

In summary, CE in FP inference is similar to CE in causal inference using observational studies. However, unlike in causal inference, the exchangeability pertains to a single outcome between C and U in FP inference.

Model-based vs. design-based methods for FP inferences

Dr. Saegusa effectively outlined the fundamental differences of model-based vs. design-based methods for FP inference. Model-based methods treat the outcome as the random variable while the design-based methods consider the selection (into the sample) indicator as random (with the outcome constant). In this paper, a set of design-based pseudoweights were constructed for the nonprobability sample to estimate FP mean under the CE assumption of $E\{y \mid b(\mathbf{x}), C\} = E\{y \mid b(\mathbf{x}), U\}$, where the expectation $E(.)$ is with respect to two levels of randomness of 1) random realization of the FP from a superpopulation, and 2) random self-selection into C from the finite population U. Dr. Saegusa further indicated that “a clear definition of y in C and/or U is desired”. The results, however, for obtaining unbiased estimation of the FP mean, apply as long as the FP is a random realization of a superpopulation. Only the existence of a distribution function with appropriate finite moments of variables is needed for the superpopulation. There is no need to specify a specific form of the parametric model; see for example Graubard and Korn (2002) for further details.

2. Response to comments by Dr. Jae-Kwang Kim and Ms. Yonghyun Kwon

In the following, I address the thoughtful discussion by Dr. Kim and Ms. Kwon in which they first presented a comprehensive framework established for categorizing existing approaches for estimating propensity scores (PS) into conditional and unconditional approaches. They further conducted simulation studies comparing different estimators, and I am glad to see our proposed ABS estimator worked well in their simulations.

In the conditional approaches two phases are involved with the first phase of sampling $S_1 = A \cup B$ from U and the second phase of sampling B from S_1. The model parameters ϕ of the conditional inclusion probability for the second phase $P(i \in S_2 \mid i \in S_1) = p(x_i;\phi)$ is estimated by

$$\hat{\phi} = \arg\max \sum_{i \in S_1} \left[\delta_i \log(p(x_i;\phi)) + (1-\delta_i)\log(1-p(x_i;\phi)) \right],$$

where $\delta_i = I(i \in B)$ is the indicator function of the event $i \in B$. Under a statistical model, say logistic regression model $\operatorname{logit}\{p(x_i;\phi)\} = x_i'\phi$, the $\hat{\phi}$ can be obtained by solving for ϕ

$$\sum_{i \in B} \left(1 - p(x_i;\phi)\right) x_i - \sum_{i \in A \text{ and } i \notin B} p(x_i;\phi)\, x_i = 0.$$

Note the overlapped units that are selected into both samples of A and B need to be identified and removed from the sample A for the second summation above. There was an accidental omission of " and $i \notin B$" under the second summation in the discussion. Based on the estimate $\hat{\phi}$, Dr. Kim and Ms. Kwon proposed the CPS pseudoweight for the i^{th} unit in B, given by

$$\hat{w}_i^{(B)} = 1 + w_i^{(A)} \left(\frac{1}{p(x_i;\hat{\phi})} - 1 \right),$$

where $w_i^{(A)}$ is often unknown and estimated under a parametric model in practice.

In the unconditional approaches, only one step was involved. The conditional maximum likelihood estimator of ϕ was estimated from the combined sample $S_1 = A \cup B$. Same as the CPS estimator, the proposed unconditional propensity score (UCPS) approach is also based on the assumption that the units that belong to the intersection of A and B can be identified.

The proposed CPS and UCPS were evaluated by simulation studies, considering varying sample sizes in sample B selected using stratified simple random sampling (SSRS) with one categorical stratification variable. This design, although simple, is clever. It aligns with the true underlying PS model for all methods considered, ensuring a fair comparison. To further evaluate the performance of the proposed estimates, we expanded the simulation studies by including an additional estimator under the same SSRS sampling design but with varying sampling weights. Recall the population size is $N =$ 5,000, sample A size $n_A =$ 250, and varying sample B sizes $n_B =$ 250 and 2,500. We consider the three estimators that have the smallest root mean squared errors (RMSE) in Table 4.1 of their discussion: WBS, ABS and CPS. The UCPS performs similarly to CPS, and therefore not considered. Recall that WBS refers to the adjusted logistic propensity estimator, proposed by Wang et al. (2021). In the same paper, the authors also proposed the scaled WBS estimator, denoted by sWBS, where the scaled weights are the value of one for sample B units and $n_s w_i^{(A)} \big/ \sum_{i \in S_A} w_i^{(A)}$ for unit i in sample A. Sampling fractions vary within sampling strata. In stratum1, $n_{B1} = f_1 n_B$ samples are selected by simple random sampling. In stratum 2, $n_{B2} = (1 - f_1)\, n_B$ samples are selected by simple random sampling. The value of f_1 is varied as 0.7, 0.8, and 0.9 to produce different values of the coefficient of variation of the SSRS sampling weights (CVWT). In the PS analysis, we consider two models: M1) the main effects of (x_1, x_2, x_3) and their pairwise interaction effects; M2) the main effects (x_{1c}, x_2, x_3) and their pairwise interaction effects, where $x_{1c} = I\{x_1 < 0\}$, the indicator function of the event $x_1 < 0$ where x_1 are generated from a $N(0,1)$. Note that M2 aligns with the SSRS design while M1 is misspecified by including the continuous variable x_1 in the PS analysis.

Four observations are made: 1) All the four estimators are approximately unbiased under the true PS model. 2) ABS and CPS perform similarly for a small sample size of $n_B =$ 250 under both models. 3) When the sample size is large $n_B =$ 2,500, CPS consistently has smallest SE and RMSE under the true model. These results are as expected, given that there is a large percentage of overlapped units in both samples. Therefore, efficiency is gained by the CPS method, which assumes that the overlapped units can be identified. 4) Under the misspecified PS model, sWBS consistently has the smallest bias, especially when CVWT is large. In contrast, CPS has the largest bias and SE when CVWT and n_B is large. The biasness and the loss of efficiency from CPS can be attributed to the misspecified modeling of $P(i \in B \mid i \in A \cup B)$, the limited sample size by removing overlapped units (~50%) from sample A, and the variable sampling weights. CPS is sensitive to model misspecification, especially when n_B and CVWT are large.

In summary, under true PS model, ABS and CPS perform similarly when n_B is small; when n_B is large, CPS estimator is more efficient due to increasing number of units that are selected and identified in both samples A and B. Under the misspecified PS model, sWBS (Wang et al., 2021) overperformed or was comparable to the other estimators. Effects of various misspecified PS models or scalers on the performance of sWBS require further investigation. Secondly, the estimators ABS, WBS, sWBS, as well as CPS, are developed without assuming that the overlapped units in both samples are negligible. For large sample size $n_B =$ 2,500, the sampling rate for sample B, $n_B / N =$ 50%, is non-negligible. All estimators, as shown in Table 1, are approximately unbiased under the true PS model, which empirically proves that all the four methods do not require the assumption that the overlapped units in both samples are negligible. Finally, it is of practical importance for the reader to be aware that the CPS estimator requires the identification of overlap units. This may not be feasible in many situations. For example, in the NIH SARS-CoV-2 Seropositivity Study discussed in my paper, this identifying information was not collected.

Table 1
Bias, standard error, and root mean square error (× 100) under SSRS with varying CV of sample weights (CVWT) after 5,000 repetitions.

	Correctly Specified PS model (x_{1c})						**Misspecified PS model (x_1)**					
	n_B = 250			**n_B = 2,500**			**n_B = 250**			**n_B = 2,500**		
	BIAS	**SE**	**RMSE**	**BIAS**	**SE**	**RMSE**	**BIAS**	**SE**	**RMSE**	**BIAS**	**SE**	**RMSE**
CVWT = 0.44												
Mean C	-3.25	2.98	4.41	-3.33	0.94	3.46	4.84	2.87	5.63	4.74	0.93	4.83
WBS	0.04	3.60	3.60	-0.04	1.42	1.42	0.51	3.06	3.10	0.23	1.54	1.56
sWBS	0.02	3.56	3.56	-0.05	1.41	1.41	0.33	3.05	3.07	0.18	1.62	1.63
ABS	0.02	3.54	3.54	-0.04	1.38	1.38	0.56	2.98	3.04	0.56	1.45	1.55
CPS	0.01	3.53	3.53	0.01	1.20	1.20	0.55	2.99	3.04	0.12	1.36	1.36
CVWT = 0.75												
Mean C	-4.97	2.89	5.76	-4.99	0.92	5.07	7.10	2.97	7.70	7.10	0.95	7.16
WBS	-0.03	4.16	4.16	0	1.56	1.56	1.03	3.18	3.34	0.47	1.55	1.62
sWBS	-0.08	4.09	4.09	-0.03	1.55	1.55	0.39	3.23	3.25	0.18	1.64	1.65
ABS	-0.08	4.08	4.08	-0.02	1.52	1.52	1.13	3.15	3.34	1.11	1.52	1.88
CPS	-0.08	4.07	4.07	0.05	1.34	1.34	1.10	3.16	3.34	-0.41	1.64	1.69
CVWT = 1.33												
Mean C	-6.58	2.88	7.19	-6.66	0.90	6.72	9.49	3.07	9.98	9.45	0.97	9.50
WBS	0.11	5.65	5.65	0.00	1.91	1.91	2.74	3.49	4.44	1.39	1.67	2.17
sWBS	0.06	5.54	5.54	-0.03	1.89	1.89	1.07	3.78	3.93	-0.05	1.85	1.85
ABS	0.03	5.49	5.49	-0.04	1.86	1.86	2.60	3.50	4.36	1.94	1.69	2.58
CPS	0.03	5.49	5.49	0.03	1.71	1.71	2.54	3.54	4.36	-3.00	3.10	4.31

References

Graubard, B.I., and Korn, E.L. (2002). Inference for superpopulation parameters using sample surveys. *Statistical Science*, 17(1), 73-96.

Rubin, D.B. (2007). The design versus the analysis of observational studies for causal effects: Parallels with the design of randomized trials. *Stat Med.*, 26(1), 20-36. Doi: 10.1002/sim.2739. PMID: 17072897.

Wang, L., Valliant, R. and Li, Y. (2021). Adjusted logistic propensity weighting methods for population inference using nonprobability volunteer-based epidemiologic cohorts. *Stat Med.*, 40(24), 5237-5250. Doi: 10.1002/sim.9122. Epub 2021 Jul 5. PMID: 34219260; PMCID: PMC8526388.

Survey Methodology, June 2024
Vol. 50, No. 1, pp. 77-106
Statistics Canada, Catalogue No. 12-001-X

Handling non-probability samples through inverse probability weighting with an application to Statistics Canada's crowdsourcing data

Jean-François Beaumont, Keven Bosa, Andrew Brennan, Joanne Charlebois and Kenneth Chu[1]

Abstract

Non-probability samples are being increasingly explored in National Statistical Offices as an alternative to probability samples. However, it is well known that the use of a non-probability sample alone may produce estimates with significant bias due to the unknown nature of the underlying selection mechanism. Bias reduction can be achieved by integrating data from the non-probability sample with data from a probability sample provided that both samples contain auxiliary variables in common. We focus on inverse probability weighting methods, which involve modelling the probability of participation in the non-probability sample. First, we consider the logistic model along with pseudo maximum likelihood estimation. We propose a variable selection procedure based on a modified Akaike Information Criterion (AIC) that properly accounts for the data structure and the probability sampling design. We also propose a simple rank-based method of forming homogeneous post-strata. Then, we extend the Classification and Regression Trees (CART) algorithm to this data integration scenario, while again properly accounting for the probability sampling design. A bootstrap variance estimator is proposed that reflects two sources of variability: the probability sampling design and the participation model. Our methods are illustrated using Statistics Canada's crowdsourcing and survey data.

Key Words: Akaike Information Criterion; Classification and Regression Trees; Logistic model; Participation probability; Statistical data integration; Variable selection.

1. Introduction

Non-probability samples are being increasingly explored at Statistics Canada and in other statistical agencies around the world. Indeed, Statistics Canada has recently conducted several non-probability surveys to evaluate the impacts of the COVID-19 pandemic on different aspects of life of the Canadian population. Data of these non-probability surveys were collected from visitors of Statistics Canada's website who responded voluntarily to an online survey questionnaire. The main motivation for considering this non-probability approach, called crowdsourcing at Statistics Canada, over probability surveys is the significant reduction in time and cost that can be achieved in the production of survey statistics. Another important advantage is the non-intrusive nature of crowdsourcing since participation is made on a voluntary basis. However, it is well known that the use of a non-probability sample alone, such as a crowdsourcing sample, may produce estimates with significant bias due to the unknown nature of the underlying selection (or participation) mechanism. To reduce this participation bias, data from a non-probability sample can be combined with data from a probability sample, ideally a large one. Estimation methods that combine data from probability and non-probability samples fall under the area of statistical data integration.

We consider the data integration scenario for which the variables of interest are available only in the non-probability sample. However, a vector of auxiliary variables is observed in both samples and used to

[1] Jean-François Beaumont, Keven Bosa, Andrew Brennan, Joanne Charlebois and Kenneth Chu, Statistics Canada, 150 Tunney's Pasture Driveway, Ottawa, Ontario, K1A 0T6. E-mail: jean-francois.beaumont@statcan.gc.ca, keven.bosa@statcan.gc.ca, andrew.brennan@statcan.gc.ca, joanne.charlebois@statcan.gc.ca and kenneth.chu@statcan.gc.ca.

reduce bias. A possible approach to inference under this scenario relies on a model for the variables of interest along with the assumption that the non-probability sample is not informative with respect to the model. The prediction approach for finite populations (e.g., Royall, 1970; Valliant, Dorfman and Royall, 2000) is one possible avenue for data integration. If a linear model between the variables of interest and the auxiliary variables holds, it can be implemented by weighting the non-probability sample through calibration on known population totals or totals estimated from the probability survey (e.g., Elliott and Valliant, 2017; Valliant, 2020). Another model-based method is statistical matching (see Yang, Kim and Hwang, 2021, for a recent reference). It consists of imputing the missing values of the variables of interest in the probability sample using non-probability sample data. The method is called sample matching (e.g., Rivers, 2007) when donor imputation is used to fill in the missing values. The prediction approach with estimated totals and statistical matching lead to identical estimators under linear models with error variance linearly related to auxiliary variables (Beaumont, 2020). Since both methods rely on a model for the variables of interest, they may become impractical when there are multiple variables of interest as a model needs to be determined and validated for each of them.

An alternative approach to inference relies on a model for the participation indicator rather than a model for the variables of interest. This approach is more appealing when there are multiple variables of interest as there is only one participation indicator, and thus only one model to choose and validate. Estimates are obtained by weighting each participant in the non-probability sample by the inverse of its estimated participation probability. This is often called inverse probability weighting or propensity score weighting in the literature. We focus on this approach. If the values of the auxiliary variables are observed for the entire population, the problem is basically identical to weighting for survey nonresponse, and nonresponse weighting methods can be applied directly to weight the non-probability sample.

In general, the auxiliary variables are observed only for the participants in the non-probability sample. Chen, Li and Wu (2020) proposed a simple and attractive method to address this issue. It requires the auxiliary variables to be also observed in a probability sample and assumes that the logistic function is used to model the participation probability. An alternative to Chen, Li and Wu (2020) consists of creating a pooled sample from the probability and nonprobability samples and modelling the participation indicator under the assumption that there is no overlap between the two samples (e.g., Lee, 2006; Valliant and Dever, 2011; and Ferri-Garcia and Rueda, 2018). Chen, Li and Wu (2020) noted that this pooling method leads to a biased estimator of the participation probability. However, Beaumont (2020) pointed out that it yields estimated participation probabilities approximately equivalent to those of Chen, Li and Wu (2020) when all the participation probabilities are small and the probability sample is properly weighted. Wang, Valliant and Li (2021) proposed an extension of the pooling method to account for a non-negligible overlap between the probability and non-probability samples. Elliott and Valliant (2017) proposed another inverse probability weighting method based on the pooled sample. It also assumes no overlap between both samples and requires the probability survey weights to be available in the non-probability sample. Recent reviews of statistical data integration methods are given in Beaumont (2020), Lohr (2021), Rao (2021), Valliant (2020), Wu (2022) and Yang and Kim (2020).

The choice of auxiliary variables is key for bias reduction. They should ideally be related to both the participation indicator and the variables of interest. Chen, Li and Wu (2020) supposed that the auxiliary variables were given. In practice, there may be a number of auxiliary variables available in both samples, often categorical, and it may not be obvious to determine the relevant ones along with proper interactions. Variable selection tools could be useful but need to be adapted to the data integration scenario considered in this paper. In particular, they need to account for the sampling design used to select the probability sample and for any adjustments to the design weights, such as nonresponse and calibration adjustments. We propose a stepwise selection procedure that achieves this goal. It is based on a modification of the Akaike Information Criterion (AIC) similar to the one Lumley and Scott (2015) developed for the estimation of model parameters from probability survey data. The Least Absolute Shrinkage and Selection Operator (LASSO) is an alternative that is considered by Bahamyirou and Schnitzer (2021). This technique usually involves cross-validation for the determination of the penalty parameter. The development of cross-validation methods that handle a combination of a probability and non-probability sample, and that properly account for the probability sampling design, requires further research.

The logistic model may sometimes produce a few estimated probabilities that are very small leading to very large weights and potentially unstable estimates. A common solution to this problem in the context of survey nonresponse is to create homogeneous groups and weight each respondent (participant) in a given group by the inverse of the estimated response (participation) rate in the group. The resulting weights possess a calibration property (see Section 3.3), which tends to limit the magnitude of the largest weights. The creation of homogeneous groups also provides some robustness to model misspecifications, as illustrated by Haziza and Lesage (2016) in the context of survey nonresponse.

A possible avenue to the creation of homogeneous groups is to adapt the Classification and Regression Trees (CART) algorithm, developed by Breiman, Friedman, Olshen and Stone (1984), to the data integration scenario studied in this paper. A nice advantage of tree-based methods is that auxiliary variables and their interactions are chosen automatically. Chu and Beaumont (2019) developed an algorithm for growing a tree that accounts for the survey weights. They called the algorithm “nppCART” because it integrates data from both a non-probability and probability sample. Pruning is an important aspect of CART that is used to avoid overfitting and to improve the efficiency of the resulting estimates. Pruning is often based on cross-validation techniques but, as pointed out above, these techniques have yet to be extended to the data integration scenario studied in this paper. Instead, we consider a modification of the AIC, similar to Lumley and Scott (2015), that properly accounts for the probability sampling design and any design weight adjustments, and use it to develop a pruning procedure.

In Section 2, we introduce the data integration problem along with some notation. The estimation of participation probabilities is discussed in Sections 3 and 4. In Section 3, we consider more specifically the logistic model and describe our proposed variable selection procedure as well as a simple rank-based method, called the Frank method, for the creation of homogeneous groups. In Section 4, we describe nppCART and our proposed pruning procedure. Bootstrap estimation of the variance of our estimators is discussed in Section 5. An empirical evaluation of our methods using real data is shown in Section 6. The last section contains some concluding remarks.

2. Data integration scenario

Let us consider the estimation of the population total $\theta = \sum_{k \in U} y_k$, where U is the set of population units and y_k is the value of a variable of interest y for population unit k. We assume that y_k is observed without error in a non-probability sample $s_{\text{NP}} \subset U$. Along with y_k, a vector of auxiliary variables $\mathbf{x}_k$ is also observed for each unit $k \in s_{\text{NP}}$. The indicator of participation in the non-probability sample is denoted by δ_k, i.e., $\delta_k = 1$, if $k \in s_{\text{NP}}$, and $\delta_k = 0$, otherwise. A probability sample s_P, drawn using some probability sampling design, is also available. The auxiliary variables $\mathbf{x}_k$ are observed for each unit $k \in s_P$, but the variable of interest y_k and the participation indicator δ_k are missing in the probability sample.

The objective is to estimate θ under the above data integration scenario, i.e., using the y values observed in the non-probability sample along with the **x** values observed in both samples. Inverse probability weighting involves modelling the participation probability $p_k = \Pr(\delta_k = 1 \mid \mathbf{x}_k)$, which is assumed to be strictly greater than 0. The estimator of θ under this approach is $\hat{\theta}_{\text{NP}} = \sum_{k \in s_{\text{NP}}} \hat{w}_k^{\text{NP}} y_k$, where $\hat{w}_k^{\text{NP}} = \hat{p}_k^{-1}$ is the non-probability survey weight, also called the pseudo survey weight, of participant k, and $\hat{p}_k$ is a consistent estimator of p_k. A critical assumption for the validity of this approach is that the participation mechanism is not informative, i.e., $\Pr(\delta_k = 1 \mid \mathbf{x}_k, y_k) = \Pr(\delta_k = 1 \mid \mathbf{x}_k)$. The availability of auxiliary variables associated with both δ_k and y_k is key to making this assumption plausible and reducing the participation bias.

The non-probability survey weight $\hat{w}_k^{\text{NP}}$ can then be calibrated (e.g., Deville and Särndal, 1992) to achieve greater efficiency gains as well as a double robustness property (e.g., Chen, Li and Wu, 2020; Valliant, 2020). Calibration of the non-probability survey weight $\hat{w}_k^{\text{NP}}$ may be particularly efficient when auxiliary variables strongly predictive of y_k are available, which were excluded from the modelling of p_k. We focus next on the modelling and estimation of the participation probability p_k.

3. Estimation of the participation probability using a logistic model

The most common model for the participation probability $p_k = \Pr(\delta_k = 1 \mid \mathbf{x}_k)$ is the logistic model $p_k(\boldsymbol{\alpha}) = \left[1 + \exp(-\mathbf{x}_k' \boldsymbol{\alpha})\right]^{-1}$, where $\boldsymbol{\alpha}$ is a vector of unknown model parameters. Assuming $\mathbf{x}_k$ is observed for all $k \in U$, and δ_k are mutually independent, an estimator of $\boldsymbol{\alpha}$ can be found by solving the unbiased maximum likelihood estimating equation:

$$\mathbf{U}(\boldsymbol{\alpha}) = \sum_{k \in U} \left[\delta_k - p_k(\boldsymbol{\alpha})\right] \mathbf{x}_k = \sum_{k \in s_{\text{NP}}} \mathbf{x}_k - \sum_{k \in U} p_k(\boldsymbol{\alpha})\, \mathbf{x}_k = \mathbf{0}. \tag{3.1}$$

The resulting maximum likelihood estimator is denoted by $\tilde{\boldsymbol{\alpha}}$ and satisfies $\mathbf{U}(\tilde{\boldsymbol{\alpha}}) = \mathbf{0}$. The estimated participation probability is denoted by $\tilde{p}_k = p_k(\tilde{\boldsymbol{\alpha}})$.

The estimating equation (3.1) cannot be used when the vector of auxiliary variables $\mathbf{x}_k$ is only observed for $k \in s_{\text{NP}}$ and missing for $k \in U - s_{\text{NP}}$. Chen, Li and Wu (2020) proposed to estimate $\sum_{k \in U} p_k(\boldsymbol{\alpha}) \mathbf{x}_k$ in (3.1) using a probability survey. The resulting pseudo maximum likelihood estimating equation is

$$\hat{\mathbf{U}}(\boldsymbol{\alpha}) = \sum\nolimits_{k \in s_{\mathrm{NP}}} \mathbf{x}_k - \sum\nolimits_{k \in s_P} w_k p_k(\boldsymbol{\alpha})\,\mathbf{x}_k = \mathbf{0}, \tag{3.2}$$

where w_k is the probability survey weight for unit $k \in s_P$. For simplicity, we assume in our theoretical developments that $w_k = \pi_k^{-1}$, where π_k is the probability that population unit k is selected in s_P. This weight ensures that $E_d\,[\hat{\mathbf{U}}(\boldsymbol{\alpha})] = \mathbf{U}(\boldsymbol{\alpha})$, where the subscript d indicates that the expectation is taken with respect to the probability sampling design. As a result, the estimating equation (3.2) is unbiased with respect to both the participation model and the probability sampling design. In practice, the survey weight w_k is often obtained after adjusting the basic design weight, π_k^{-1}, for nonresponse and calibration. The estimating equation (3.2) requires knowing the vector $\mathbf{x}_k$ for all $k \in s_{\mathrm{NP}}$ and all $k \in s_P$ but not for all $k \in U$. Its solution yields the pseudo maximum likelihood estimator $\hat{\boldsymbol{\alpha}}$, which satisfies $\hat{\mathbf{U}}(\hat{\boldsymbol{\alpha}}) = \mathbf{0}$. The resulting estimated participation probability is denoted by $\hat{p}_k = p_k(\hat{\boldsymbol{\alpha}})$. Note that the estimating equation (3.2) may not have a solution. This is more likely to occur when n^{NP}/N is large and the probability sample is small (see Beaumont, 2020). This was not an issue in our experimentations since n^{NP}/N was smaller than 1%. Beaumont (2020) argued that the occurrence of inexistent solutions may be reduced by replacing the logistic model with the exponential model.

Chen, Li and Wu (2020) considered the case where the auxiliary variables are given. In practice, it may be necessary to choose relevant auxiliary variables and their interactions among a large set of candidate auxiliary variables. In the applications we have experimented with so far, the candidate auxiliary variables are often categorical (e.g., education, marital status, etc.). Blindly crossing all these variables may lead to a huge number of groups with many small groups, even empty. This was our motivation for finding methods that could select relevant auxiliary variables and their interactions.

We consider a stepwise selection procedure that attempts to minimize a modified version of the AIC, which properly accounts for the probability sampling design used to draw s_P. The justification for this modified AIC is provided in Section 3.1, and our selection procedure is described in Section 3.2. Section 3.3 considers an important special case of the logistic model: the homogeneous group model. In Section 3.4, a simple rank-based method for creating homogeneous groups is proposed. Finally, in Section 3.5, the recent method of Wang, Valliant and Li (2021) is discussed and contrasted with the method of Chen, Li and Wu (2020).

3.1 A modified AIC for the logistic model that accounts for the probability sampling design

Let us first consider the case where $\mathbf{x}_k$ is known for all the population units $k \in U$. Assuming δ_k are mutually independent, we can write the log likelihood function as

$$\begin{aligned} l(\boldsymbol{\alpha}) &= \sum\nolimits_{k \in U} \delta_k \log[p_k(\boldsymbol{\alpha})] + (1-\delta_k)\log[1-p_k(\boldsymbol{\alpha})] \\ &= \sum\nolimits_{k \in s_{\mathrm{NP}}} \log\left[\frac{p_k(\boldsymbol{\alpha})}{1-p_k(\boldsymbol{\alpha})}\right] + \sum\nolimits_{k \in U} \log[1-p_k(\boldsymbol{\alpha})]. \end{aligned}$$

Let us define $l_0(\boldsymbol{\alpha}) = E_m[l(\boldsymbol{\alpha})],$ where the subscript m indicates that the expectation is taken with respect to the true unknown participation model. The maximum likelihood estimator $\tilde{\boldsymbol{\alpha}}$ maximizes $l(\boldsymbol{\alpha})$ and we denote by $\boldsymbol{\alpha}_0,$ the value of $\boldsymbol{\alpha}$ that maximizes $l_0(\boldsymbol{\alpha}).$ Under regularity conditions, the maximum likelihood estimator $\tilde{\boldsymbol{\alpha}}$ is consistent for $\boldsymbol{\alpha}_0$ under the model, i.e., $\sqrt{N}\left(\tilde{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0\right) = O_p(1),$ where N is the population size.

The AIC is an estimator of $-2E_m[l_0(\tilde{\boldsymbol{\alpha}})].$ It is well known that a consistent estimator of $-2E_m[l_0(\tilde{\boldsymbol{\alpha}})]$ is

$$\text{AIC} = -2l(\tilde{\boldsymbol{\alpha}}) + 2q, \tag{3.3}$$

where q is the number of model parameters (or the number of auxiliary variables). Equation (3.3) is the original AIC expression and the most widespread in practice.

Let us now consider the case where $\mathbf{x}_k$ is known only for $k \in s_{\text{NP}}$ and $k \in s_P.$ Chen, Li and Wu (2020) proposed the pseudo log likelihood function

$$\hat{l}(\boldsymbol{\alpha}) = \sum\nolimits_{k \in s_{\text{NP}}} \log\left[\frac{p_k(\boldsymbol{\alpha})}{1 - p_k(\boldsymbol{\alpha})}\right] + \sum\nolimits_{k \in s_P} w_k \log[1 - p_k(\boldsymbol{\alpha})]. \tag{3.4}$$

Using $w_k = \pi_k^{-1}$ ensures that $E_d[\hat{l}(\boldsymbol{\alpha})] = l(\boldsymbol{\alpha})$ and $E_{md}[\hat{l}(\boldsymbol{\alpha})] = l_0(\boldsymbol{\alpha}).$ Under regularity conditions, the pseudo maximum likelihood estimator $\hat{\boldsymbol{\alpha}},$ which maximizes $\hat{l}(\boldsymbol{\alpha})$ in (3.4), is consistent for $\boldsymbol{\alpha}_0$ under both the model and the sampling design, i.e., $\sqrt{n^P}\,(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0) = O_p(1),$ where n^P is the size of the probability sample.

Under pseudo maximum likelihood estimation, the AIC can be defined as an estimator of

$$-2E_{md}[l_0(\hat{\boldsymbol{\alpha}})] = -2E_{md}[\hat{l}(\hat{\boldsymbol{\alpha}})] + 2E_{md}[\hat{l}(\hat{\boldsymbol{\alpha}}) - l_0(\hat{\boldsymbol{\alpha}})].$$

In Appendix 1, we provide a sketch of the proof that

$$E_{md}[\hat{l}(\hat{\boldsymbol{\alpha}}) - l_0(\hat{\boldsymbol{\alpha}})] \approx q + \text{tr}\left[E_m\{\text{var}_d[\hat{\mathbf{U}}(\boldsymbol{\alpha}_0)]\}[-\mathbf{H}_0(\boldsymbol{\alpha}_0)]^{-1}\right], \tag{3.5}$$

where the function $\hat{\mathbf{U}}(\boldsymbol{\alpha}) = \partial\hat{l}(\boldsymbol{\alpha})/\partial\boldsymbol{\alpha}$ is given in (3.2) for the logistic model, and $\mathbf{H}_0(\boldsymbol{\alpha}) = \partial^2 l_0(\boldsymbol{\alpha})/\partial\boldsymbol{\alpha}\,\partial\boldsymbol{\alpha}'.$ Our derivations follow closely those of Lumley and Scott (2015). From (3.5) and (A.3) in Appendix 1, a consistent estimator of $-2E_{md}[l_0(\hat{\boldsymbol{\alpha}})]$ is

$$\text{AIC} = -2\hat{l}(\hat{\boldsymbol{\alpha}}) + 2q + 2\text{tr}\left\{\hat{\mathbf{v}}_d[\hat{\mathbf{U}}(\boldsymbol{\alpha}_0)][-\hat{\mathbf{H}}(\hat{\boldsymbol{\alpha}})]^{-1}\right\}, \tag{3.6}$$

where $\hat{\mathbf{v}}_d[\hat{\mathbf{U}}(\boldsymbol{\alpha}_0)]$ is any design-consistent estimator of $\text{var}_d[\hat{\mathbf{U}}(\boldsymbol{\alpha}_0)]$ and $\hat{\mathbf{H}}(\boldsymbol{\alpha}) = \partial^2\hat{l}(\boldsymbol{\alpha})/\partial\boldsymbol{\alpha}\,\partial\boldsymbol{\alpha}'.$ For the logistic model,

$$\hat{\mathbf{H}}(\boldsymbol{\alpha}) = -\sum\nolimits_{k \in s_P} w_k p_k(\boldsymbol{\alpha})[1 - p_k(\boldsymbol{\alpha})]\,\mathbf{x}_k\mathbf{x}_k'. \tag{3.7}$$

The AIC expression (3.6) is similar to the one given in Lumley and Scott (2015) but they omitted the term $2q.$ This term is negligible compared with the third term on the right-hand side of (3.6) when the sampling fraction n^P/N is negligible. However, the term $2q$ may not always be negligible compared with the third term of (3.6), even when n^P/N is small. This would tend to occur when the participation probabilities $p_k(\boldsymbol{\alpha})$ are small, which is typically the case of online volunteer-based surveys like Statistics Canada's crowdsourcing surveys. Therefore, the term $2q$ should generally not be neglected unless the non-probability sample size is significantly larger than the probability sample size. Another reason for keeping $2q$ in the expression (3.6) is that it reduces to the standard AIC expression (3.3) when the probability sample is a census. The last term on the right-hand side of (3.6) can thus be interpreted as a penalty for using a probability sample instead of a complete census in the estimating equation (3.2). The smaller the probability sample, the larger the effect of the penalty on the AIC (3.6).

3.2 Stepwise selection of auxiliary variables and pairwise interactions

In the empirical Section 6, we use a stepwise procedure based on the AIC (3.6) to select auxiliary variables (main effects) and pairwise interactions. Our procedure starts with the naïve model, which only includes the intercept. At each step of the procedure, a variable (main effect or pairwise interaction) is either included in the model or, if it was previously included, removed from the model. The inclusion or removal of the variable that yields the largest reduction of the AIC (3.6) is selected. An interaction is only eligible for inclusion when both main effects have already been selected, and a main effect is only eligible for removal when it is not supporting any interaction. The procedure stops when no variable can be added or removed from the model, i.e., no further reduction of the AIC (3.6) is possible.

One issue with the selection of auxiliary variables in a participation model is that it ignores the relationships between auxiliary variables and the variables of interest. As a result, an auxiliary variable that would be weakly associated with participation but strongly associated with some of the variables of interest could be discarded from the final participation model. This could have a negative effect on the bias reduction of the estimator $\hat{\theta}_{\mathrm{NP}}$ of the finite population parameter θ. It is thus advisable to consider variable selection methods that lean towards overfitting, such as the AIC, to reduce the risk of omitting a relevant auxiliary variable. Moderate overfitting may better control for bias at the expense of a possible increase in variance. Our intent is to avoid gross overfitting so as to stabilize $\hat{\theta}_{\mathrm{NP}}$. As pointed out in Section 2, the above variable selection issue can also be dealt with by calibrating inverse probability weights $\hat{w}_k^{\mathrm{NP}}$ using calibration variables that are predictive of the variables of interest.

3.3 The homogeneous group model

Consider a partition of the population U into G groups, U_g, $g=1,...,G$, and let $s_{\mathrm{NP},g}$ and $s_{P,g}$ be the sets of units $k \in U_g$ that fall in the non-probability and probability samples, respectively. In the homogeneous group model, the participation probability is assumed to be constant for all units $k \in U_g$, i.e., $p_k \equiv p_g$, $k \in U_g$, $g=1,...,G$. The homogeneous group model can be viewed as a special case of the logistic model with $q=G$, $\boldsymbol{\alpha}'=(\alpha_1,\ldots,\alpha_g,\ldots,\alpha_G)$ and $\mathbf{x}_k'=(x_{1k},\ldots,x_{gk},\ldots,x_{Gk})$, where x_{gk} is a binary variable that equals 1 if $k \in U_g$, and that equals 0, otherwise. Therefore, for a unit $k \in U_g$, $p_k(\boldsymbol{\alpha})=p(\alpha_g)\equiv p_g = [1+\exp(-\alpha_g)]^{-1}$, and thus $\alpha_g = \log[p_g/(1-p_g)]$. For this model, the pseudo log likelihood function (3.4) reduces to

$$\hat{l}(\boldsymbol{\alpha}) = \sum_{g=1}^{G} n_g^{\text{NP}} \log\left[\frac{p(\alpha_g)}{1-p(\alpha_g)}\right] + \hat{N}_g \log[1-p(\alpha_g)], \tag{3.8}$$

where n_g^{NP} is the size of $s_{\text{NP},g}$ and $\hat{N}_g = \sum_{k \in s_{P,g}} w_k$ is the estimated population size in group g obtained from the probability sample. The pseudo maximum likelihood estimator $\hat{\boldsymbol{\alpha}}' = (\hat{\alpha}_1, \ldots, \hat{\alpha}_g, \ldots, \hat{\alpha}_G),$ which maximizes $\hat{l}(\boldsymbol{\alpha})$ in (3.8), is such that $\hat{\alpha}_g = \log[\hat{p}_g / (1-\hat{p}_g)],\ g = 1, \ldots, G,$ where

$$\hat{p}_g = \frac{n_g^{\text{NP}}}{\hat{N}_g}. \tag{3.9}$$

From (3.8), we can write $\hat{l}(\hat{\boldsymbol{\alpha}})$ as

$$\hat{l}(\hat{\boldsymbol{\alpha}}) = \sum_{g=1}^{G} \hat{N}_g [\hat{p}_g \log(\hat{p}_g) + (1-\hat{p}_g)\log(1-\hat{p}_g)]. \tag{3.10}$$

For the homogeneous group model, the estimating function $\hat{\mathbf{U}}(\boldsymbol{\alpha})$ in (3.2) reduces to $[\hat{\mathbf{U}}(\boldsymbol{\alpha})]' = [\hat{U}_1(\alpha_1), \ldots, \hat{U}_g(\alpha_g), \ldots, \hat{U}_G(\alpha_G)],$ where

$$\hat{U}_g(\alpha_g) = n_g^{\text{NP}} - \hat{N}_g p(\alpha_g). \tag{3.11}$$

Also, from (3.7), the matrix $\hat{\mathbf{H}}(\hat{\boldsymbol{\alpha}})$ reduces to a diagonal matrix with the g^{th} element on the diagonal given by

$$\hat{H}_g(\hat{\alpha}_g) = -\hat{N}_g \hat{p}_g (1-\hat{p}_g). \tag{3.12}$$

Let $\boldsymbol{\alpha}_0' = (\alpha_{0,1}, \ldots, \alpha_{0,g}, \ldots, \alpha_{0,G}).$ Using (3.11) and (3.12), the AIC (3.6) becomes

$$\text{AIC} = -2\hat{l}(\hat{\boldsymbol{\alpha}}) + 2G + 2\sum_{g=1}^{G} \frac{\hat{v}_d[\hat{U}_g(\alpha_{0,g})]}{\hat{N}_g \hat{p}_g (1-\hat{p}_g)}, \tag{3.13}$$

where $\hat{v}_d[\hat{U}_g(\alpha_{0,g})]$ is a design-consistent estimator of $\text{var}_d[\hat{U}_g(\alpha_{0,g})].$ Using (3.11), a consistent variance estimator is

$$\hat{v}_d[\hat{U}_g(\alpha_{0,g})] = \hat{p}_g^2 \hat{v}_d(\hat{N}_g), \tag{3.14}$$

where $\hat{v}_d(\hat{N}_g)$ is a design-consistent estimator of $\text{var}_d(\hat{N}_g).$ Using (3.14), the AIC (3.13) can be rewritten as

$$\text{AIC} = -2\hat{l}(\hat{\boldsymbol{\alpha}}) + 2G + 2\sum_{g=1}^{G} n_g^{\text{NP}} \frac{[\text{cv}_d(\hat{N}_g)]^2}{1-\hat{p}_g}, \tag{3.15}$$

where $\text{cv}_d(\hat{N}_g) = \sqrt{\hat{v}_d(\hat{N}_g)} \big/ \hat{N}_g$ is the estimated coefficient of variation of $\hat{N}_g$. Again, the last term on the right-hand side of (3.13) or (3.15) can be interpreted as a penalty for estimating the unknown population sizes $N_g,\ g = 1, \ldots, G,$ using a probability sample.

Using (3.9), we obtain the non-probability survey weight of a unit $k \in s_{\mathrm{NP},g}$ as

$$\hat{w}_k^{\mathrm{NP}} = \hat{p}_k^{-1} = \frac{\hat{N}_g}{n_g^{\mathrm{NP}}}. \tag{3.16}$$

The non-probability survey weight (3.16) shows the importance of avoiding groups for which n_g^{NP} is very small, even zero, so as to reduce the occurrence of extreme weights. Using (3.16), the inverse probability weighted estimator of the population total θ can be written as

$$\hat{\theta}_{\mathrm{NP}} = \sum\nolimits_{k \in s_{\mathrm{NP}}} \hat{w}_k^{\mathrm{NP}} y_k = \sum_{g=1}^{G} \hat{N}_g \bar{y}_g^{\mathrm{NP}}, \tag{3.17}$$

where $\bar{y}_g^{\mathrm{NP}} = \sum\nolimits_{k \in s_{\mathrm{NP},g}} y_k \Big/ n_g^{\mathrm{NP}}$ is the average of variable of interest y over units in $s_{\mathrm{NP},g}$. The estimator (3.17) is simply a post-stratified estimator and satisfies the calibration equations $\sum\nolimits_{k \in s_{\mathrm{NP},g}} \hat{w}_k^{\mathrm{NP}} = \hat{N}_g$, $g = 1, \ldots, G$. The groups (post-strata) are constructed to be homogeneous with respect to the participation indicator. If they are also homogeneous with respect to the variable of interest then the post-stratified estimator (3.17) has a double robustness property (e.g., see Chen, Li and Wu, 2020; and Valliant, 2020).

We have assumed so far that the group membership is pre-determined for every population unit. In practice, homogeneous groups are often defined after observing sample data. There are several methods of constructing sample-dependent homogeneous groups. In Section 3.4, we propose a simple rank-based method that partitions the non-probability sample with respect to estimated participation probabilities from a logistic model. An extension of CART, nppCART, is described in Section 4. Once the non-probability and probability samples have been partitioned into sample-dependent groups, weights can be computed using (3.16) as if the group memberships were fixed.

3.4 A rank-based method for creating homogeneous groups

The first step of this method consists of estimating participation probabilities using a logistic model (with or without stepwise selection). We denote by $\hat{p}_k^{\text{logistic}} = p_k(\hat{\boldsymbol{\alpha}})$ these estimated participation probabilities, which are computed for each $k \in s_{\mathrm{NP}}$ and $k \in s_P$. The idea is then to form G groups that are homogeneous with respect to $\hat{p}_k^{\text{logistic}}$ so as to make the homogeneous group model plausible. Once the groups are formed, the estimated probabilities $\hat{p}_k^{\text{logistic}}$ are discarded and the non-probability survey weights are computed using (3.16).

There are many methods for partitioning s_{NP} into homogeneous groups. A simple and popular method is to form groups with an equal number of participants (e.g., Eltinge and Yansaneh, 1997, formed groups with an equal number of sample units in the context of survey nonresponse). This method is equivalent to determining group boundaries from equal-width intervals in the range of r_k, $k \in s_{\mathrm{NP}}$, where r_k is the rank of $\hat{p}_k^{\text{logistic}}$. We propose below a generalization of this method that retains the simplicity of assigning units based on their rank, but allows some flexibility so that the classes do not need to be equal-sized.

Rather than making equal-width bins in the range of r_k, we propose to form G equal-width bins in the range of $f(r_k)$, a monotone function of the rank r_k. We call it the Frank method. All the non-probability sample units that fall in a given bin are assigned to the same group. Any non-linear function f would thus make smaller groups (fewer units) where the slope is steeper and larger groups where the slope is flatter. We propose the function

$$f(r_k) = \log\left(1 + a\frac{r_k}{n^{\text{NP}}}\right),$$

$k \in s_{\text{NP}}$, where n^{NP} is the size of the non-probability sample and a is a non-negative pre-specified constant that determines the degree of non-linearity. This function is concave down, with a larger slope and smaller groups for the lower-ranked units. The constant a determines the size of this effect, with a large value (e.g., $a = 100$) providing groups that are more unequal in size. The limit as a approaches 0 from above renders this function linear and so returns the equal-sized groups. The rank can be defined in ascending order of $\hat{p}_k^{\text{logistic}}$ ($r_k = 1$ for the smallest $\hat{p}_k^{\text{logistic}}$, $r_k = 2$ for the second smallest $\hat{p}_k^{\text{logistic}}$, etc.), in which case the units with smaller estimated probabilities will be in the smaller groups, or in descending order of $\hat{p}_k^{\text{logistic}}$ ($r_k = 1$ for the largest $\hat{p}_k^{\text{logistic}}$, $r_k = 2$ for the second largest $\hat{p}_k^{\text{logistic}}$, etc.), in which case the units with larger estimated probabilities will be in the smaller groups. The Frank method is somewhat similar to forming equal-width groups but with the groups bunched toward one end or the other, depending on whether $\hat{p}_k^{\text{logistic}}$ are sorted in ascending or descending order. Figure A.1(A) in Appendix 2 illustrates the Frank method for $a = 10$, $G = 15$ and n^{NP} = 31,415, which is the size of the non-probability sample used in our empirical study in Section 6.

Once the non-probability sample has been partitioned into groups, each probability sample unit must then be assigned to one of the groups. Because the function f is monotone, each group contains non-probability sample units with values of $\hat{p}_k^{\text{logistic}}$ within a certain interval, and the intervals of any two different groups do not overlap so that the groups can be sorted based on their average value of $\hat{p}_k^{\text{logistic}}$. The boundary between any two consecutive groups is taken as the midpoint between the largest $\hat{p}_k^{\text{logistic}}$ from the group with the smaller average and the smallest $\hat{p}_k^{\text{logistic}}$ from the other group. Once all the boundaries have been determined, each probability sample unit $k \in s_P$ is assigned to the group with boundaries that cover $\hat{p}_k^{\text{logistic}}$.

The application of the Frank method requires determining suitable values of a and G as well as sorting $\hat{p}_k^{\text{logistic}}$, $k \in s_{\text{NP}}$, in ascending or descending order before computing the ranks r_k. Each possible choice leads to a different set of groups. We propose to determine the values of a and G, and the sorting order, by looking at different options and choosing the one that yields the smallest value of the AIC (3.15). This is investigated empirically in Section 6.3.

3.5 Adjusted logistic propensity weighting

As pointed out in the introduction, Wang, Valliant and Li (2021) proposed an extension of the pooling method to account for a non-negligible overlap between the probability and non-probability samples. The justification of their method, called Adjusted Logistic Propensity (ALP) weighting, is not based on a true likelihood approach, but still yields an *md*-unbiased estimating equation given by

$$\hat{\mathbf{U}}^{\text{ALP}}(\boldsymbol{\alpha}) = \sum\nolimits_{k \in s_{\text{NP}}} \frac{1}{1 + p_k^{\text{ALP}}(\boldsymbol{\alpha})}\,\mathbf{x}_k - \sum\nolimits_{k \in s_P} w_k \frac{p_k^{\text{ALP}}(\boldsymbol{\alpha})}{1 + p_k^{\text{ALP}}(\boldsymbol{\alpha})}\,\mathbf{x}_k = \mathbf{0}, \tag{3.18}$$

where $p_k^{\mathrm{ALP}}(\boldsymbol{\alpha}) = \exp(\mathbf{x}_k'\boldsymbol{\alpha})$. The estimating equation (3.18) is not equivalent to (3.2). However, if all the participation probabilities are small, both estimating equations should yield similar estimates of the participation probabilities.

An important difference between Wang, Valliant and Li (2021) and Chen, Li and Wu (2020) is the choice of the participation model. Chen, Li and Wu (2020) modelled the participation probability using a logistic function whereas Wang, Valliant and Li (2021) considered an exponential function. The logistic model is more natural as it ensures that estimated participation probabilities are always within the $(0,1)$ interval. This is to be contrasted with the exponential model, which may produce estimated probabilities greater than 1. Wang, Valliant and Li (2021) conducted a simulation study to evaluate their method. Their results show that (3.18) yields estimates of population means that are more robust to model failure than (3.2). This robustness could be explained by the use of the exponential model.

For the homogeneous group model, we have seen in Section 3.3 that the solution of (3.2) yields $p_k(\hat{\boldsymbol{\alpha}}) = \hat{p}_g = n_g^{\mathrm{NP}} \big/ \hat{N}_g$, for every unit $k \in U_g$. It is straightforward to show that the solution of (3.18) for the homogeneous group model also yields $p_k^{\mathrm{ALP}}(\hat{\boldsymbol{\alpha}}) = \hat{p}_g = n_g^{\mathrm{NP}} \big/ \hat{N}_g$, for every unit $k \in U_g$. The equivalence between (3.2) and (3.18) for the homogeneous group model suggests that, in general, the two methods may produce similar estimates of θ, particularly when estimated probabilities are used only for the purpose of creating homogeneous groups (e.g., using the Frank method described in Section 3.4).

Wang, Valliant and Li (2021) also proposed a scaled version of their ALP method. Although the scaled estimating equation is not *md*-unbiased anymore, the authors showed its effectiveness in a simulation study for the estimation of population means. We tested the ALP method, including its scaled version, in our empirical experiments. The resulting estimates (not reported) were close to the pseudo maximum likelihood estimates of Chen, Li and Wu (2020), particularly after creating homogeneous groups. This observation is not surprising considering that the non-probability sample size is smaller than 1% of the population size in our experiments and that the estimated participation probabilities tend to be quite small. A thorough comparison of ALP and pseudo maximum likelihood estimation is left for future research.

One of the objectives of this paper was to develop a variable selection procedure applicable to the data integration scenario described in Section 2. Wang, Valliant and Li (2021) did not tackle the problem of variable selection. An AIC based on Lumley and Scott (2015) is not appropriate with ALP (or its scaled version) because the underlying estimating equation is not justified through a true likelihood approach. However, if ALP were preferable in a given context, variable selection could first be based on the pseudo likelihood method of Chen, Li and Wu (2020) and then ALP could be applied using the selected auxiliary variables.

4. Estimation of the participation probability using nppCART

The CART tree-growing procedure, developed by Breiman, Friedman, Olshen and Stone (1984), is a recursive binary partitioning algorithm that minimizes a certain objective function. For a binary dependent variable such as δ_k, a suitable objective function is the entropy impurity. For a given partition, U_g, $g = 1, \ldots, G$, the entropy impurity is given by

$$I = -\sum_{g=1}^{G} \frac{N_g}{N} [\tilde{p}_g \log(\tilde{p}_g) + (1-\tilde{p}_g) \log(1-\tilde{p}_g)],$$

where N_g is the size of U_g, $N = \sum_{g=1}^{G} N_g$ and $\tilde{p}_g = n_g^{\text{NP}} / N_g$. The entropy impurity cannot be computed when N_g is unknown. We propose to replace N_g with the survey-weighted estimator $\hat{N}_g$. This yields the computable objective function

$$\hat{I} = -\sum_{g=1}^{G} \frac{\hat{N}_g}{\hat{N}} [\hat{p}_g \log(\hat{p}_g) + (1-\hat{p}_g) \log(1-\hat{p}_g)], \tag{4.1}$$

where $\hat{p}_g$ is given in (3.9) and $\hat{N} = \sum_{g=1}^{G} \hat{N}_g$. The estimated entropy impurity (4.1) is proportional to the pseudo log likelihood function (3.10) under the homogeneous group model since $\hat{I} = -\hat{l}(\hat{\boldsymbol{\alpha}}) / \hat{N}$.

The recursive binary partitioning algorithm starts by examining all the possible splits of the non-probability sample s_{NP} into two groups. A split is any binary partition of s_{NP} based on the categories or numerical values of one of the candidate auxiliary variables. For instance, a split could be "SEX = male" and "SEX = female" or "AGE < 25" and "AGE ≥ 25". For each split of s_{NP}, the probability sample s_P is also split using the same binary partition. A split is said to be inadmissible and is rejected if it satisfies any of the following three stopping criteria:

i) $n_g^{\text{NP}} < C_{\text{NP}}$, for $g = 1$ or $g = 2$, where $C_{\text{NP}} \geq 1$ is a pre-determined constant specifying the minimum number of participants in a group;

ii) $n_g^{\text{NP}} \geq \hat{N}_g$, for $g = 1$ or $g = 2$;

iii) $n_g^P < C_P$, for $g = 1$ or $g = 2$, where n_g^P is the size of $s_{P,g}$ and $C_P \geq 1$ is a pre-determined constant specifying the minimum number of probability sample units in a group.

Then, the estimated entropy impurity (4.1) with $G = 2$ is computed for each admissible split, and the best of those admissible splits, i.e., the one that has the smallest value of (4.1), is selected to form the first two groups. If all the splits are inadmissible or the best split does not decrease the objective function (4.1) then partitioning into two groups is not done.

After the determination of the first two initial groups, the same splitting operation is repeated for each of the two groups, and so on and so forth, layer by layer, until all the groups cannot be split further based on the stopping criteria. We say that this process results in a fully grown tree although it is a slight abuse of language as there are stopping criteria that limit its growth. The above procedure, the earlier version of which was called nppCART by Chu and Beaumont (2019), is essentially identical to the original CART algorithm, except for the use of the estimated entropy (4.1) and the three stopping criteria above. The stopping criterion (i) ensures that the non-probability survey weight $\hat{w}_k^{\text{NP}}$ in (3.16) does not become extreme. The stopping criterion (ii) ensures that the estimated probability $\hat{p}_g$ is always smaller than 1. The last criterion is added to ensure that the estimator $\hat{N}_g$ is not too unstable.

Chu and Beaumont (2019) developed an R program that implements the nppCART algorithm. They showed in a simulation study that this algorithm was effective for reducing the participation bias although the resulting post-stratified estimator (3.17) had a variance somewhat larger than its competitors. This instability might be explained by overfitting, i.e., the creation of too many groups. The usual recommendation to avoid overfitting is to prune the tree after it has been grown. Pruning is usually applied in two steps. In the first step, a finite sequence of nested subtrees of decreasing size and increasing impurity is determined, starting with the fully grown tree that has the maximum number of groups and ending with the degenerate subtree that contains only one group. In the second step, the best of these nested subtrees is selected, often through K-fold cross-validation. This pruning approach is equivalent to penalizing the objective function with an additive penalty term defined as the product of a positive penalty parameter and the number of groups. Cross-validation is then typically used to determine an optimal value for the penalty parameter. Greater detail on pruning can be found in Breiman, Friedman, Olshen and Stone (1984); see also Izenman (2008, Chapter 9). In the context of survey nonresponse, classification and regression trees have been explored by Phipps and Toth (2012) and Lohr, Hsu and Montaquila (2015).

However, as pointed out in the introduction, classical cross-validation methods cannot be directly applied to the data integration scenario studied in this paper, and this topic requires further research. As an alternative to cross-validation for the selection of the best subtree, among a set of nested subtrees of decreasing size and increasing impurity, we propose to choose the subtree that minimizes the AIC (3.15). This AIC takes the probability sampling design into account through the estimation of the design variance of $\hat{N}_g$ (see Section 5). This variance could be readily estimated in our experiments in Section 6 using available bootstrap weights. Similar to variable selection, discussed in Section 3.2, pruning is intended to avoid gross overfitting so as to stabilize $\hat{\theta}_{\text{NP}}$.

5. Bootstrap variance estimation

It is not enough to produce inverse probability weighted estimates of finite population parameters; it is also important to provide users with indicators of the quality of those estimates. We propose a bootstrap procedure to estimate the variance of inverse probability weighted estimators with a focus on the post-stratified estimator (3.17). The variance may be useful but has some limitations since it is derived under the assumption that the participation model is correctly specified and that the inverse probability weighted estimators are unbiased. The absence of bias depends critically on the availability and proper choice of auxiliary variables so as to make the non-informative participation assumption reasonable. Although some amount of bias seems unavoidable in practice, the computation of variance estimates may nonetheless provide some useful information for comparison and evaluation purposes, as illustrated in Section 6.

The bootstrap variance estimator that we propose accounts for two sources of variability: the probability sampling design and the participation model. We suppose that B bootstrap weights $w_k^{(b)}, b = 1, \ldots, B,$ are available for each unit $k \in s_P,$ and that these bootstrap weights properly capture the variability due to the probability sampling design. For instance, we assume that these bootstrap weights can be used to obtain a design-consistent estimator of $\text{var}_d(\hat{N}_g)$ as

$$\hat{v}_d^{\text{boot}}(\hat{N}_g) = \frac{1}{B} \sum_{b=1}^{B} (\hat{N}_g^{(b)} - \hat{N}_g)^2, \tag{5.1}$$

where $\hat{N}_g^{(b)} = \sum_{k \in s_{P,g}} w_k^{(b)}$ is the b^{th} bootstrap replicate of $\hat{N}_g$. The Rao, Wu and Yue (1992) bootstrap weights are often used in social surveys conducted by Statistics Canada. They are applicable for stratified multistage designs when the first-stage sampling fractions are small and can incorporate weight adjustments, such as nonresponse adjustments and calibration. Beaumont and Émond (2022) proposed an extension of the method that removes the requirement of small first-stage sampling fractions.

The unknown participation mechanism is modelled as a Poisson sampling design, where population units are assumed to participate independently of one another with probability p_k, $k \in U$. For Poisson sampling, Beaumont and Patak (2012) pointed out that valid bootstrap weights for sample units $k \in s_{\text{NP}}$ can be written as $p_k^{-1} a_k^{(b)}$, $b = 1, \ldots, B$, provided that the bootstrap factors $a_k^{(b)}$ are generated independently of one another using a distribution that is not too heavily skewed with a mean of one and a variance of $1 - p_k$. For a non-probability sample, the true participation probability p_k is unknown but can be replaced with a consistent estimator $\hat{p}_k$. Following Beaumont and Émond (2022), who studied bootstrap under survey nonresponse, we thus suggest generating the bootstrap factors $a_k^{(b)}$, $k \in s_{\text{NP}}$ and $b = 1, \ldots, B$, independently of one another using the gamma distribution with a mean of one and a variance of $1 - \hat{p}_k$. The choice of the gamma distribution is to ensure non-negative bootstrap factors $a_k^{(b)}$.

The bootstrap estimator of the variance of the inverse probability weighted estimator $\hat{\theta}_{\text{NP}}$, $\text{var}_{md}(\hat{\theta}_{\text{NP}})$, is given by

$$\hat{v}_{md}^{\text{boot}}(\hat{\theta}_{\text{NP}}) = \frac{1}{B} \sum_{b=1}^{B} (\hat{\theta}_{\text{NP}}^{(b)} - \hat{\theta}_{\text{NP}})^2, \tag{5.2}$$

where $\hat{\theta}_{\text{NP}}^{(b)}$ is the b^{th} bootstrap replicate of $\hat{\theta}_{\text{NP}}$. Assuming the logistic model is used with fixed auxiliary variables, the b^{th} bootstrap replicate of $\hat{\theta}_{\text{NP}} = \sum_{k \in s_{\text{NP}}} \hat{w}_k^{\text{NP}} y_k$, with $\hat{w}_k^{\text{NP}} = [p_k(\hat{\boldsymbol{\alpha}})]^{-1}$, is $\hat{\theta}_{\text{NP}}^{(b)} = \sum_{k \in s_{\text{NP}}} \hat{w}_k^{\text{NP},(b)} y_k$, where $\hat{w}_k^{\text{NP},(b)} = a_k^{(b)} / p_k(\hat{\boldsymbol{\alpha}}^{(b)})$, and $\hat{\boldsymbol{\alpha}}^{(b)}$ is the solution of the b^{th} bootstrap replicate of estimating equation (3.2):

$$\hat{\mathbf{U}}^{(b)}(\boldsymbol{\alpha}) = \sum_{k \in s_{\text{NP}}} a_k^{(b)} \mathbf{x}_k - \sum_{k \in s_P} w_k^{(b)} p_k(\boldsymbol{\alpha}) \, \mathbf{x}_k = \mathbf{0}.$$

Assuming now that the homogeneous group model is used, the b^{th} bootstrap replicate of the post-stratified estimator (3.17) can be written as

$$\hat{\theta}_{\text{NP}}^{(b)} = \sum_{k \in s_{\text{NP}}} \hat{w}_k^{\text{NP},(b)} y_k = \sum_{g=1}^{G} \hat{N}_g^{(b)} \bar{y}_g^{\text{NP},(b)}, \tag{5.3}$$

where $\hat{w}_k^{\text{NP},(b)} = a_k^{(b)} \hat{N}_g^{(b)} / n_g^{\text{NP},(b)}$, for $k \in s_{\text{NP},g}$, $n_g^{\text{NP},(b)} = \sum_{k \in s_{\text{NP},g}} a_k^{(b)}$ and $\bar{y}_g^{\text{NP},(b)} = \sum_{k \in s_{\text{NP},g}} a_k^{(b)} y_k / n_g^{\text{NP},(b)}$. The bootstrap replicate (5.3) is valid provided that the homogeneous groups are fixed. This simplification is often made when estimating the variance of estimators adjusted for survey nonresponse, even when the homogeneous groups are determined adaptively from the observed sample data. In our context, it would not be straightforward to develop a bootstrap procedure that correctly accounts for variable selection or pruning. In particular, a double bootstrap might be required if the design variance estimators involved in the AIC (3.6) or (3.15) were obtained through bootstrap weights. Treating auxiliary variables or homogeneous groups as fixed, when they are not, should tend to underestimate the variance $\text{var}_{md}(\hat{\theta}_{\text{NP}})$. Although the magnitude of the underestimation is expected to be small to moderate, further research is needed on this topic.

6. Empirical evaluation of methods using real data

We evaluated and compared inverse probability weighting methods, discussed in Sections 3 and 4, using real data. In Section 6.1, we present the three data sources used in our investigations. Methods are described in Section 6.2 and results are given in Sections 6.3 and 6.4.

6.1 Data sources and variables

After the beginning of the COVID-19 lockdown in March 2020, Statistics Canada conducted a series of crowdsourcing surveys to respond to urgent information needs about the life of the Canadian population. Each crowdsourcing survey collected data from visitors of Statistics Canada's website who responded voluntarily to a short online questionnaire. Renaud and Beaumont (2020) provide greater detail on crowdsourcing experiments conducted by Statistics Canada.

We investigated the use of the Labour Force Survey (LFS) as a means of reducing the participation bias of crowdsourcing estimates. Except for the Census, the LFS is the most important social probability survey conducted by Statistics Canada with a sample containing around 56,000 selected households each month. Data are collected for all eligible persons within responding households. The household response rate was around 90% before the pandemic but fell to around 70% in June 2020. In our empirical study, we used data of the June 2020 LFS sample, which contains responses for 87,779 persons. The LFS is based on a stratified multistage design and a regression composite estimator (see Gambino, Kennedy and Singh, 2001). Rao, Wu and Yue (1992) bootstrap weights are produced and made available to users for variance estimation.

In parallel to crowdsourcing experiments, Statistics Canada also started a series of probability web panel surveys: the Canadian Perspective Survey Series (CPSS). The CPSS sample is obtained from previous LFS respondents. The June 2020 CPSS initial probability sample was relatively large with over 30,000 selected persons but the overall recruitment/response rate was quite low at around 15%; this resulted in 4,209 respondents in June 2020. Greater detail on the CPSS can be found in Baribeau (2020).

In June 2020, participants from previous crowdsourcing experiments were also randomly chosen and sent the same questionnaire as CPSS respondents; 31,415 participants responded to the questionnaire. This allowed for a comparison of estimates from this crowdsourcing non-probability sample with those from the CPSS probability sample.

Table 6.1 shows naïve crowdsourcing estimates and CPSS estimates for nine selected proportions. For the first two proportions, LFS estimates are also available and very close to the corresponding CPSS estimates. This is not unexpected as nonresponse in the CPSS is adjusted using education and employment status. Both probability surveys show large differences with naïve crowdsourcing estimates for these two proportions. The following five proportions also show significant differences between naïve crowdsourcing and CPSS estimates whereas estimates from both sources are similar for the last two proportions.

Table 6.1
Proportions of interest.

Proportion	Description	Naïve crowdsourcing estimate	CPSS estimate	LFS estimate
θ_1	Proportion of people having a university degree.	64.5%	30.6%	30.2%
θ_2	Proportion of people who worked at a job or business during the reference week.	65.4%	50.1%	50.3%
θ_3	Proportion of people whose usual place of work is a fixed location outside the home.	50.2%	40.2%	-
θ_4	Proportion of people who worked most of their hours at home during the reference week.	45.6%	19.3%	-
θ_5	Proportion of people who report having "more than enough" income to meet their household needs.	32.1%	15.9%	-
θ_6	Proportion of people who are "very likely" to get COVID-19 vaccine when available.	74.2%	57.3%	-
θ_7	Proportion of people who are "very concerned" about the health risk posed by gathering in large groups.	70.0%	54.4%	-
θ_8	Proportion of people who "fear being a target for putting others at risk" because they do not always wear a mask in public.	9.9%	9.8%	-
θ_9	Proportion of people who report ordering the same amount of take-out food as before.	45.6%	46.2%	-

In a first step, we used June 2020 LFS data to reduce the participation bias of naïve crowdsourcing estimates using inverse probability weighting methods discussed in Sections 3 and 4. The candidate auxiliary variables available in both the crowdsourcing and LFS samples were: age group (13 levels), sex (2 levels), economic region (56 levels), education (8 levels), immigration status (3 levels), household size (6 levels), marital status (6 levels) and employment status (3 levels). Greater detail on these eight auxiliary variables is given in Appendix 3. Then, we used non-probability survey weights to compute adjusted crowdsourcing estimates for the nine proportions defined in Table 6.1 and compared them to those obtained using the CPSS probability sample alone. These results are provided in Section 6.3. Note that a proportion is defined as $\theta = N^{-1}\sum_{k\in U} y_k$, where y_k is a binary variable of interest, and is estimated by $\hat{\theta}_{\text{NP}} = \sum_{k\in s_{\text{NP}}} \hat{w}_k^{\text{NP}} y_k \Big/ \sum_{k\in s_{\text{NP}}} \hat{w}_k^{\text{NP}}$. For the first two proportions in Table 6.1, the variable of interest y_k can be derived from auxiliary variables. We thus expect weighting methods to successfully remove the participation bias for these proportions.

In a second step, we obtained adjusted crowdsourcing estimates using June 2020 CPSS data instead of LFS data with the same candidate auxiliary variables as above. Our objective was to evaluate the effect on bias reduction of using a smaller probability sample. These results are provided in Section 6.4.

6.2 Methods

We investigated the eight methods described in Table 6.2 below. For methods 3, 5 and 6, which involve a logistic model with the stepwise selection procedure described in Section 3.2, all main effects and pairwise interactions were considered as candidate variables to be included or removed from the model. For these methods, the estimator $\hat{\mathbf{v}}_d\,[\hat{\mathbf{U}}(\boldsymbol{\alpha}_0)]$, required to compute the AIC (3.6), was obtained using bootstrap weights as

$$\hat{\mathbf{v}}_d[\hat{\mathbf{U}}(\boldsymbol{\alpha}_0)] = \frac{1}{B}\sum_{b=1}^{B}\,[\hat{\mathbf{U}}^{*(b)}(\hat{\boldsymbol{\alpha}})]\,[\hat{\mathbf{U}}^{*(b)}(\hat{\boldsymbol{\alpha}})]',$$

where

$$\hat{\mathbf{U}}^{*(b)}(\hat{\boldsymbol{\alpha}}) = \sum\nolimits_{k\in s_{\mathrm{NP}}}\mathbf{x}_k - \sum\nolimits_{k\in s_P} w_k^{(b)}\,p_k(\hat{\boldsymbol{\alpha}})\,\mathbf{x}_k.$$

For methods 4, 5, 6 and 8, the estimator $\hat{v}_d(\hat{N}_g)$, required to compute the AIC (3.15), is obtained from (5.1). For methods, 6, 7 and 8, which use nppCART, we set $C_{\mathrm{NP}} = 5$ and $C_P = 5$ in the stopping criteria (i) and (iii) given in Section 4.

Table 6.2
Description of methods.

Method	Model	Stepwise selection	Homogeneous groups	Description
1	Intercept	-	-	Naïve logistic model with only the intercept (or homogeneous group model with only one group).
2	Logistic	-	-	Logistic model including all main effects but no interaction.
3	Logistic	Yes	-	Logistic model with stepwise selection of main effects and pairwise interactions by minimizing the AIC (3.6).
4	Logistic	-	Frank	Method 2 followed by creation of homogeneous groups using the Frank method, described in Section 3.4, with sorting in ascending order, $a = 10$ and the number of groups roughly minimizing the AIC (3.15).
5	Logistic	Yes	Frank	Method 3 followed by creation of homogeneous groups using the Frank method, described in Section 3.4, with sorting in ascending order, $a = 10$ and the number of groups roughly minimizing the AIC (3.15).
6	Logistic	Yes	nppCART with pruning	Method 3 followed by creation of homogeneous groups using nppCART with pruning minimizing the AIC (3.15); only one auxiliary variable is provided to nppCART: the estimated participation probability from the logistic model.
7	-	-	nppCART without pruning	nppCART based on all candidate auxiliary variables without pruning.
8	-	-	nppCART with pruning	nppCART based on all candidate auxiliary variables with pruning minimizing the AIC (3.15).

6.3 Results when integrating crowdsourcing data with the LFS probability sample

Stepwise selection results for the logistic model

Using the LFS as the probability sample, our stepwise selection procedure described in Section 3.2 resulted in the selection of all main effects along with 15 pairwise interactions for a total of 395 model parameters. Six main effects entered the model before any interaction in the following order: education, economic region, immigration status, sex, age group and household size. Together, they accounted for more than 95% of the total AIC reduction (difference between AIC of methods 1 and 3). The variable education alone accounted for more than 40% of the total AIC reduction. For these data, it thus appears that interactions are not as important as the main effects to reduce the AIC. This suggests that a model including all the main effects but no interaction might be reasonable.

Comparisons of AIC values

Table 6.3 shows values of the Relative AIC (RAIC) for the eight methods described in Table 6.2. The Relative AIC is defined as

$$\mathrm{RAIC} = \frac{\mathrm{AIC}_0 - \mathrm{AIC}}{\mathrm{AIC}_0} \times 100\%,$$

where AIC_0 is the value of the AIC (3.6) for the naïve model containing only the intercept. For methods 1, 2 and 3, the RAIC is computed using the AIC (3.6) whereas it is computed using the AIC (3.15) for methods 4 to 8 assuming the groups are fixed. The RAIC can be interpreted similarly to the coefficient of determination in linear regression: it is 0 for the naïve model, it increases as the AIC decreases, and it is always smaller than 1. However, it can take negative values unlike the coefficient of determination. A model that has a larger RAIC than a competitor suggests that its auxiliary variables are better predictors of participation. Table 6.3 also shows the number of model parameters q or the number of groups $G; q$ is shown for methods 1, 2 and 3, and G is shown for methods 4 to 8.

Table 6.3
RAIC values in percentage.

Method	**Model**	**Stepwise selection**	**Homogeneous groups**	**RAIC (%)**	***q* or *G***	**Proportion (%) of AIC from the 1st term**	**Proportion (%) of AIC from the 2nd term**	**Proportion (%) of AIC from the 3rd term**
1	Intercept	-	-	0	1	100.00	0.00	0.00
2	Logistic	-	-	10.7	90	99.90	0.04	0.06
3	Logistic	Yes	-	11.1	395	99.59	0.18	0.23
4	Logistic	-	Frank	10.7	100	99.89	0.05	0.07
5	Logistic	Yes	Frank	11.3	100	99.88	0.05	0.07
6	Logistic	Yes	nppCART with pruning	12.2	1,276	97.99	0.59	1.42
7	-	-	nppCART without pruning	11.9	3,165	96.23	1.45	2.33
8	-	-	nppCART with pruning	12.5	1,772	97.58	0.82	1.60

The RAIC varies from 10.7% to 12.5% for methods 2 to 8; thus, all these methods provide a meaningful improvement over the naïve method. Comparing methods 2 and 3, we observe that accounting for pairwise interactions yielded only a small improvement of the RAIC, as noted above. Using the Frank method to create homogeneous groups did not significantly improve the RAIC. This is an indication that the logistic model was reasonable for these data. The use of nppCART resulted in an increase of RAIC, albeit not substantial. This may indicate that nppCART has achieved some robustness. However, nppCART also resulted in a number of groups significantly larger than other methods, even after pruning. Given the AIC (3.15) assumes the groups are fixed (although they are not), this improvement of RAIC should not be over-interpreted.

Table 6.3 also shows the proportion of the AIC that comes from each of the three terms on the right-hand side of (3.6) or (3.15). Not surprisingly, the first term, $-2\hat{l}(\hat{\boldsymbol{\alpha}}),$ is the dominant component of the AIC. The relative importance of the other two terms increases with q or $G.$ Both terms have similar importance although the third term is always slightly larger than the second term. In this application, none of the terms should be omitted in the computation of the AIC.

The Frank Method

Figures A.1(B) and A.1(C) in the Appendix 2 illustrate the Frank method of creating homogeneous groups for method 5 in Table 6.2. Figure A.1(B) shows a graph of $\hat{p}_k^{\text{logistic}}$ as a function of the rank r_k for both the non-probability and probability samples. It also shows the corresponding boundaries, in terms of the ranks, for $G = 15$ and different values of $a,$ and for both sorting orders. Figure A.1(B) illustrates that the groups containing smaller values of $\hat{p}_k^{\text{logistic}}$ are under-represented in the non-probability sample, compared with the probability sample, because these units are less likely to participate. Figure A.1(B) also illustrates that sorting in ascending order produces groups that are closer to being equal-sized in the probability sample, particularly when a is large. This has the advantage of reducing the occurrence of groups that contain too few probability sample units, which could lead to unstable weights. A value of $a = 5$ or $a = 10,$ along with sorting in ascending order, seems to offer a suitable compromise for both samples.

Figure A.1(C) shows the values of the AIC (3.15) as a function of the number of groups G for a few values of a and both sorting orders. It appears that the sorting order makes a significant difference on the AIC, with lower values obtained when $\hat{p}_k^{\text{logistic}},\ k \in s_{\text{NP}},$ are sorted in ascending order. Figure A.1(C) does not show much sensitivity to the choice of a but the best values seem to occur near $a = 10.$ Notably, the optimal number of classes is near 100 in this application, much larger than the value of 5 that is often recommended (e.g., Eltinge and Yansaneh, 1997). Based on these results, we chose to sort in ascending order and used $a = 10$ and $G = 100$ when applying the Frank method with LFS data. A smaller number of groups was chosen with the CPSS data (see Section 6.4).

With these data, forming groups with an equal number of participants $(a = 0)$ was slightly inferior to $a = 10$ in terms of AIC (see Figure A.1(C)). However, both values of a led to similar estimates (results not shown).

Comparisons of estimates

Table 6.4 shows estimates and their bootstrap standard errors (in italic) for each of the nine proportions in Table 6.1 and each method described in Table 6.2. The bootstrap standard error is the square root of the bootstrap variance estimate given in (5.2). The b^{th} bootstrap replicate of the estimated proportion $\hat{\theta}_{\text{NP}} = \sum_{k \in s_{\text{NP}}} \hat{w}_k^{\text{NP}} y_k \Big/ \sum_{k \in s_{\text{NP}}} \hat{w}_k^{\text{NP}}$ is $\hat{\theta}_{\text{NP}}^{(b)} = \sum_{k \in s_{\text{NP}}} \hat{w}_k^{\text{NP},(b)} y_k \Big/ \sum_{k \in s_{\text{NP}}} \hat{w}_k^{\text{NP},(b)}.$ For methods 4 to 8, the bootstrap weights $\hat{w}_k^{\text{NP},(b)}$ are obtained under the simplification that the homogeneous groups are fixed. Bootstrap standard errors are not computed for methods 2 and 3. The CPSS estimates and their design-based standard errors are also provided for comparison purposes in the last row of Table 6.4. The CPSS estimates are believed to be less biased than adjusted crowdsourcing estimates since they are obtained from a probability survey, albeit with a small response rate (around 15%), with nonresponse weight adjustments and calibration.

From the estimates and standard errors in Table 6.4, we make the following observations:

- Methods 2 to 8 are all roughly equivalent.
- For the first seven proportions, where the naïve estimates (method 1) are significantly different from the CPSS estimates, methods 2 to 8 yield adjusted crowdsourcing estimates closer to CPSS estimates, which suggests a non-negligible bias reduction. Indeed, for the first three proportions, the adjusted crowdsourcing estimates are not markedly different from the CPSS estimates. It is not surprising for the first two proportions since the variables of interest can be derived from auxiliary variables. This observation is particularly interesting for the third proportion. For proportions 4 to 7, the bias reduction is not so spectacular, albeit not negligible; the adjusted crowdsourcing estimates lie in between the naïve and CPSS estimates.
- For the last two proportions, the naïve, adjusted crowdsourcing and CPSS estimates are all similar. A slight but not alarming discrepancy between adjusted crowdsourcing and CPSS estimates is observed for the last proportion for methods 2 and 3, which do not use homogeneous groups. Overall, it is reassuring to observe that inverse probability weighting did not introduce significant biases for the last two proportions.
- Finally, the standard errors for the naïve method are much smaller than those for the other methods. This indicates that naïve estimates are likely more stable. However, the standard error does not account for bias and should not be the main criterion for choosing an appropriate method.

Table 6.4
Estimates and standard errors (in italic) in percentage.

Method	Model	Stepwise selection	Homogeneous groups	θ_1	θ_2	θ_3	θ_4	θ_5	θ_6	θ_7	θ_8	θ_9
1	Intercept	-	-	64.5	65.4	50.2	45.6	32.1	74.2	70.0	9.9	45.6
				0.27	*0.26*	*0.27*	*0.28*	*0.27*	*0.24*	*0.26*	*0.17*	*0.28*
2	Logistic	-	-	29.7	50.2	40.4	28.0	23.5	67.9	62.4	11.4	43.5
				-	-	-	-	-	-	-	-	-
3	Logistic	Yes	-	28.9	48.2	39.8	26.6	23.3	68.1	64.1	10.2	42.3
				-	-	-	-	-	-	-	-	-
4	Logistic	-	Frank	32.4	52.1	40.6	29.5	23.5	68.0	63.5	10.7	44.9
				0.41	*0.76*	*0.70*	*0.58*	*0.60*	*0.74*	*0.78*	*0.49*	*0.77*
5	Logistic	Yes	Frank	30.8	51.4	39.8	28.5	22.4	67.9	64.0	10.3	44.4
				0.35	*0.86*	*0.78*	*0.63*	*0.59*	*0.82*	*0.89*	*0.54*	*0.87*
6	Logistic	Yes	nppCART with pruning	30.9	50.7	39.5	28.4	22.9	67.8	63.7	10.4	44.5
				0.36	*0.84*	*0.78*	*0.70*	*0.79*	*1.02*	*1.00*	*0.62*	*1.02*
7	-	-	nppCART without pruning	30.2	52.7	40.6	28.0	24.3	69.3	65.4	9.4	46.8
				0.29	*0.88*	*0.91*	*0.46*	*0.82*	*0.91*	*0.96*	*0.42*	*0.74*
8	-	-	nppCART with pruning	30.2	52.5	40.5	28.0	23.8	69.4	65.2	9.3	47.0
				0.29	*0.87*	*0.91*	*0.47*	*0.81*	*0.90*	*1.03*	*0.39*	*0.78*
CPSS estimate				**30.6**	**50.1**	**40.2**	**19.3**	**15.9**	**57.3**	**54.4**	**9.8**	**46.2**
				0.87	***1.25***	***1.14***	***0.97***	***0.87***	***1.41***	***1.33***	***0.86***	***1.42***

With these data, methods 2 to 8 performed similarly. This may be due to the large size of the LFS probability sample. In order to study the behaviour of inverse probability weighting methods when the probability sample is smaller, we replaced the LFS by the CPSS probability sample. Results for this case are discussed below.

6.4 Results when integrating crowdsourcing data with the CPSS probability sample

Stepwise selection results for the logistic model

When we used the CPSS as the probability sample, our stepwise selection procedure selected again all main effects but only 10 pairwise interactions for a total of 254 model parameters. All but one main effect entered the model before any interaction in the following order: education, household size, economic region, sex, immigration status, age group and marital status. For these data, pairwise interactions were again not as important as the main effects to reduce the AIC.

Comparisons of AIC values

Table 6.5 shows values of the RAIC for the eight methods described in Table 6.2. Comparing methods 2 and 3, we observe that accounting for pairwise interactions yielded only a small improvement of the RAIC. For these data, the creation of homogeneous groups resulted in a non-negligible increase of the RAIC. In particular, when a logistic model is used along with stepwise selection, the RAIC is 12.1 and it increases to 18.5 after forming homogeneous groups with nppCART. The use of nppCART without a logistic model (methods 7 and 8) also yielded a larger RAIC than methods 2 and 3. The effect of pruning remains negligible with these data since the RAIC of methods 7 and 8 are similar. However, pruning reduced the number of groups from 600 to 451. The replacement of the LFS sample by the CPSS sample resulted in a reduction of the number of groups for methods 4 to 8; this is not surprising since the CPSS sample size is significantly smaller than the LFS sample size.

Table 6.5 also shows the proportion of the AIC that comes from each of the three terms on the right-hand side of (3.6) or (3.15). Again, the first term, $-2\hat{l}(\hat{\boldsymbol{\alpha}})$, is the dominant component of the AIC, and the relative importance of the other two terms increases with q or G. Given the small CPSS sample size, the third term, which can be viewed as a penalty for using a probability sample instead of a census, is now relatively much larger than the second term $2q$ (or $2G$). The second term could thus be omitted, as in Lumley and Scott (2015), although there is no computational advantage of neglecting it.

Comparisons of estimates

Table 6.6 shows estimates and their bootstrap standard errors (in italic) for each of the nine proportions in Table 6.1 and each method described in Table 6.2. We make the following observations:

- For the first two proportions, the variables of interest can be derived from auxiliary variables, and we expect inverse probability weighting methods to entirely remove bias. Methods 7 and 8 (nppCART without a logistic model) basically eliminated the discrepancy between the naïve and CPSS estimates. Other methods were not so successful although method 4 (logistic model with main effects followed by the Frank method) performed relatively well.

- Method 2 appeared to over-adjust the naïve estimates for the first three proportions. Forming homogeneous group (method 4) corrected for this over-adjustment.
- Methods 2 and 3 (logistic model without homogeneous groups) were somewhat erratic. This may be explained by variable and extreme non-probability survey weights, particularly for method 3. The coefficient of variation of the non-probability survey weights is provided in Table 6.7 for each method. It is 7.5 and 39.7 for methods 2 and 3, respectively, whereas it is no greater than 5.5 for all the other methods. This shows the importance of forming homogeneous groups to reduce extreme weights. By comparison, when the LFS is used as the probability sample, the coefficient of variation of the non-probability survey weights is 4.7 and 6.3 for methods 2 and 3, respectively, and it is no greater than 4.0 for all the other methods.
- Methods that use stepwise selection tended to under-adjust when homogeneous groups were formed (methods 5 and 6), particularly for the first proportion. This was not expected given their large values of RAIC in Table 6.5. However, the RAIC only indicates the strength of the association between the auxiliary variables and participation. It does not account for the strength of the association between the auxiliary variables and variables of interest, which can affect the magnitude of participation bias and variance.
- Comparing methods 5 and 6, we observe that the creation of homogeneous groups using the Frank method and nppCART yielded similar estimates with nppCART estimates tending to be slightly closer to CPSS estimates, possibly due to the larger number of groups with nppCART.
- Pruning did not show significant improvements in our experiments since methods 7 and 8 produced similar estimates.
- Overall, nppCART with or without pruning (methods 7 and 8) appeared to be the most stable and reliable method for reducing participation bias followed closely by method 4 (logistic model with main effects only along with the Frank method).

It is interesting to observe that nppCART estimates in Table 6.6 (methods 7 and 8) were not markedly different from the corresponding estimates in Table 6.4 based on the LFS probability sample. This suggests that a small probability sample can succeed at reducing bias even though it remains preferable to use a larger probability sample. For nppCART, using the LFS as the probability sample was just slightly better than using the CPSS. For other methods, the differences were sometimes much larger and using the LFS provided better estimates. This may be an argument to favour nppCART when the probability sample size is small.

Table 6.5
RAIC values in percentage.

Method	Model	Stepwise selection	Homogeneous groups	RAIC (%)	q or G	Proportion (%) of AIC from the 1st term	Proportion (%) of AIC from the 2nd term	Proportion (%) of AIC from the 3rd term
1	Intercept	-	-	0	1	100.00	0.00	0.00
2	Logistic	-	-	11.2	90	98.45	0.04	1.50
3	Logistic	Yes	-	12.1	254	96.27	0.12	3.62
4	Logistic	-	Frank	13.4	20	98.18	0.01	1.80
5	Logistic	Yes	Frank	15.9	16	99.35	0.01	0.64
6	Logistic	Yes	nppCART with pruning	18.5	384	96.43	0.19	3.38
7	-	-	nppCART without pruning	14.3	600	95.93	0.28	3.78
8	-	-	nppCART with pruning	14.4	451	96.27	0.21	3.51

Table 6.6
Estimates and standard errors (in italic) in percentage.

Method	Model	Stepwise selection	Homogeneous groups	θ_1	θ_2	θ_3	θ_4	θ_5	θ_6	θ_7	θ_8	θ_9
1	Intercept	-	-	64.5	65.4	50.2	45.6	32.1	74.2	70.0	9.9	45.6
				0.28	*0.28*	*0.29*	*0.29*	*0.28*	*0.25*	*0.25*	*0.17*	*0.28*
2	Logistic	-	-	21.3	44.4	34.4	24.4	22.8	69.1	61.3	10.2	44.9
				-	-	-	-	-	-	-	-	-
3	Logistic	Yes	-	29.4	43.4	28.3	29.8	27.4	78.4	71.8	10.1	27.6
				-	-	-	-	-	-	-	-	-
4	Logistic	-	Frank	34.1	50.9	39.4	30.2	25.8	70.8	66.6	9.8	45.1
				0.59	*0.61*	*0.56*	*0.51*	*0.50*	*0.55*	*0.58*	*0.36*	*0.59*
5	Logistic	Yes	Frank	43.6	54.6	41.8	34.3	27.4	71.7	67.9	9.7	44.6
				0.67	*0.54*	*0.50*	*0.55*	*0.43*	*0.44*	*0.47*	*0.30*	*0.47*
6	Logistic	Yes	nppCART with pruning	42.0	54.0	41.2	34.2	27.3	70.8	67.4	10.1	44.6
				0.81	*0.77*	*0.73*	*0.71*	*0.63*	*0.69*	*0.66*	*0.44*	*0.70*
7	-	-	nppCART without pruning	30.8	48.9	39.1	28.5	27.7	71.5	64.9	8.9	47.1
				0.98	*1.38*	*1.41*	*0.80*	*1.35*	*1.23*	*1.46*	*0.56*	*1.49*
8	-	-	nppCART with pruning	30.8	49.8	38.7	29.3	27.1	71.5	65.2	9.3	46.8
				0.98	*1.27*	*1.28*	*0.78*	*1.24*	*1.20*	*1.41*	*0.80*	*1.35*
CPSS estimate				**30.6**	**50.1**	**40.2**	**19.3**	**15.9**	**57.3**	**54.4**	**9.8**	**46.2**
				0.87	***1.25***	***1.14***	***0.97***	***0.87***	***1.41***	***1.33***	***0.86***	***1.42***

Table 6.7
Coefficients of variation of the non-probability survey weights.

Probability sample	Method 1	Method 2	Method 3	Method 4	Method 5	Method 6	Method 7	Method 8
CPSS	0	7.5	39.7	1.8	1.4	2.2	5.5	5.0
LFS	0	4.7	6.3	2.6	3.0	3.6	4.0	3.9

7. Conclusion

We extended the pseudo maximum likelihood method of Chen, Li and Wu (2020) that integrates data from a non-probability and probability sample: We developed a variable selection procedure for the logistic model and an extension of CART, nppCART. Inspired by Lumley and Scott (2015), our extensions use a modified AIC that properly accounts for the probability sampling design. In our investigations, we observed that the additional penalty term for using a probability sample instead of a census was not negligible.

Not surprisingly, our experimentations illustrated that inverse probability weighting methods can reduce participation bias, but sometimes a significant bias remains. For the large LFS probability sample, all the methods performed similarly. Significant differences between methods were observed when the smaller CPSS probability sample was used. In particular, our experimentations showed the importance of creating homogeneous groups to reduce the occurrence of extreme weights and improve the stability and robustness of estimates. For the small probability sample, accounting for pairwise interactions somewhat reduced the AIC but was generally not beneficial for the estimates. Main effects appeared more important than pairwise interactions to reduce the AIC with our data. Overall, the best method for bias reduction was nppCART followed closely by the use of a logistic model with main effects only along with the creation of homogeneous groups. However, different conclusions could potentially be drawn with smaller domains or other datasets.

It is well known that inverse probability weighted estimators may be inefficient, particularly when the variables of interest are weakly related to the weights. This can be addressed through calibration on known population totals or totals estimated from the probability sample. Calibration will be particularly efficient when auxiliary variables strongly related to the variables of interest are available and excluded from the participation model. This was not the case in our experimentations. Weight smoothing is an alternative aiming to improve the efficiency of inverse probability weighted estimators, which may be useful when such powerful calibration variables are not available. It consists of replacing the weights with predictions obtained by modelling the weights conditionally on the variables of interest. In the context of integrating non-probability and probability samples, weight smoothing was studied by Ferri-Garcia, Beaumont, Bosa, Charlebois and Chu (2021).

Tree-based methods more sophisticated than the CART algorithm, such as random forests, are available in the literature. Given the good performance of nppCART in our experimentations, it could be worthwhile to extend those methods to the data integration scenario considered in this paper and evaluate them. Further developments are needed on this topic.

There is most likely no inverse probability weighting method that is uniformly better than all the other methods. All the techniques are useful and can be part of the statistician’s toolkit. However, there is a need for the development of bias reduction indicators that would help statisticians in choosing the best method for a given non-probability and probability sample. The relative AIC and the coefficient of variation of the non-probability survey weights are two useful indicators but they do not tell the full story as they do not say anything about the strength of the association between the auxiliary variables and variables of interest. One idea that could be explored would be to use statistical matching methods with nonparametric models (e.g., random forests) for each variable of interest conditionally on the auxiliary variables. The resulting estimates would be expected to be more efficient than inverse probability weighting methods because they would be tailored to each variable of interest. In practice, this statistical matching strategy would be tedious to apply as a different model would need to be developed and validated for each estimate produced. However, a few statistical matching estimates could be computed and used to evaluate inverse probability weighting methods. We might expect that a better inverse probability weighting method would generally tend to yield estimates closer to statistical matching estimates. A possible procedure to reconcile the two methods would be to calibrate inverse probability weights so that the resulting estimates agree exactly with selected statistical matching estimates.

Appendix 1

Sketch of the proof of equation (3.5)

Using first-order Taylor expansions, we have

$$\hat{l}(\hat{\boldsymbol{\alpha}}) - l_0(\hat{\boldsymbol{\alpha}}) = \left[\hat{l}(\boldsymbol{\alpha}_0) - l_0(\boldsymbol{\alpha}_0)\right] + \left[\hat{\mathbf{U}}(\boldsymbol{\alpha}_0) - \mathbf{U}_0(\boldsymbol{\alpha}_0)\right]' \left(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0\right) + o_p\left(\frac{N}{n^P}\right) \tag{A.1}$$

and

$$\hat{\mathbf{U}}(\hat{\boldsymbol{\alpha}}) = \hat{\mathbf{U}}(\boldsymbol{\alpha}_0) + \hat{\mathbf{H}}(\boldsymbol{\alpha}_0)\left(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0\right) + o_p\left(\frac{N}{\sqrt{n^P}}\right), \tag{A.2}$$

where $\mathbf{U}_0(\boldsymbol{\alpha}) = \partial l_0(\boldsymbol{\alpha}) / \partial \boldsymbol{\alpha}$. In addition to (A.1) and (A.2), we also assume that

$$\hat{\mathbf{H}}(\boldsymbol{\alpha}) = \mathbf{H}_0(\boldsymbol{\alpha}) + o_p\left(N\right) \tag{A.3}$$

under the model and the sampling design. Noting that $\mathbf{U}_0(\boldsymbol{\alpha}_0) = \mathbf{0}$ and $\hat{\mathbf{U}}(\hat{\boldsymbol{\alpha}}) = \mathbf{0}$, we obtain from (A.1), (A.2) and (A.3),

$$\hat{l}(\hat{\boldsymbol{\alpha}}) - l_0(\hat{\boldsymbol{\alpha}}) = \left[\hat{l}(\boldsymbol{\alpha}_0) - l_0(\boldsymbol{\alpha}_0)\right] + \left(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0\right)' \left[-\mathbf{H}_0(\boldsymbol{\alpha}_0)\right]\left(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0\right) + o_p\left(\frac{N}{n^P}\right). \tag{A.4}$$

Ignoring the smaller order term and taking the expectation of both sides of (A.4) yield:

$$E_{md}\,[\hat{l}(\hat{\boldsymbol{\alpha}}) - l_0(\hat{\boldsymbol{\alpha}})] \approx \operatorname{tr}[-\mathbf{H}_0(\boldsymbol{\alpha}_0)\operatorname{var}_{md}(\hat{\boldsymbol{\alpha}})], \tag{A.5}$$

where $\operatorname{var}_{md}(\hat{\boldsymbol{\alpha}}) = E_{md}[(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0)\,(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0)']$. Using (A.2) and (A.3), and ignoring the smaller order terms, we can approximate this variance as

$$\begin{aligned}
\operatorname{var}_{md}(\hat{\boldsymbol{\alpha}}) &\approx [\mathbf{H}_0(\boldsymbol{\alpha}_0)]^{-1}\operatorname{var}_{md}[\hat{\mathbf{U}}(\boldsymbol{\alpha}_0)]\,[\mathbf{H}_0(\boldsymbol{\alpha}_0)]^{-1} \\
&= [\mathbf{H}_0(\boldsymbol{\alpha}_0)]^{-1}\left\{\operatorname{var}_m[\mathbf{U}(\boldsymbol{\alpha}_0)] + E_m\{\operatorname{var}_d[\hat{\mathbf{U}}(\boldsymbol{\alpha}_0)]\}\right\}[\mathbf{H}_0(\boldsymbol{\alpha}_0)]^{-1} \qquad \text{(A.6)} \\
&= -[\mathbf{H}_0(\boldsymbol{\alpha}_0)]^{-1} + [\mathbf{H}_0(\boldsymbol{\alpha}_0)]^{-1}E_m\{\operatorname{var}_d[\hat{\mathbf{U}}(\boldsymbol{\alpha}_0)]\}\,[\mathbf{H}_0(\boldsymbol{\alpha}_0)]^{-1},
\end{aligned}$$

where $\mathbf{U}(\boldsymbol{\alpha}) = \partial l(\boldsymbol{\alpha}) / \partial \boldsymbol{\alpha}$ is given in equation (3.1) for the logistic model. The last equation in (A.6) results from a well-known property of the Fisher information matrix $-\mathbf{H}_0(\boldsymbol{\alpha}_0)$ (assuming the true model is in the same parametric family as the postulated model). Using (A.6) in (A.5) yields result (3.5).

Appendix 2

Illustration of the Frank method

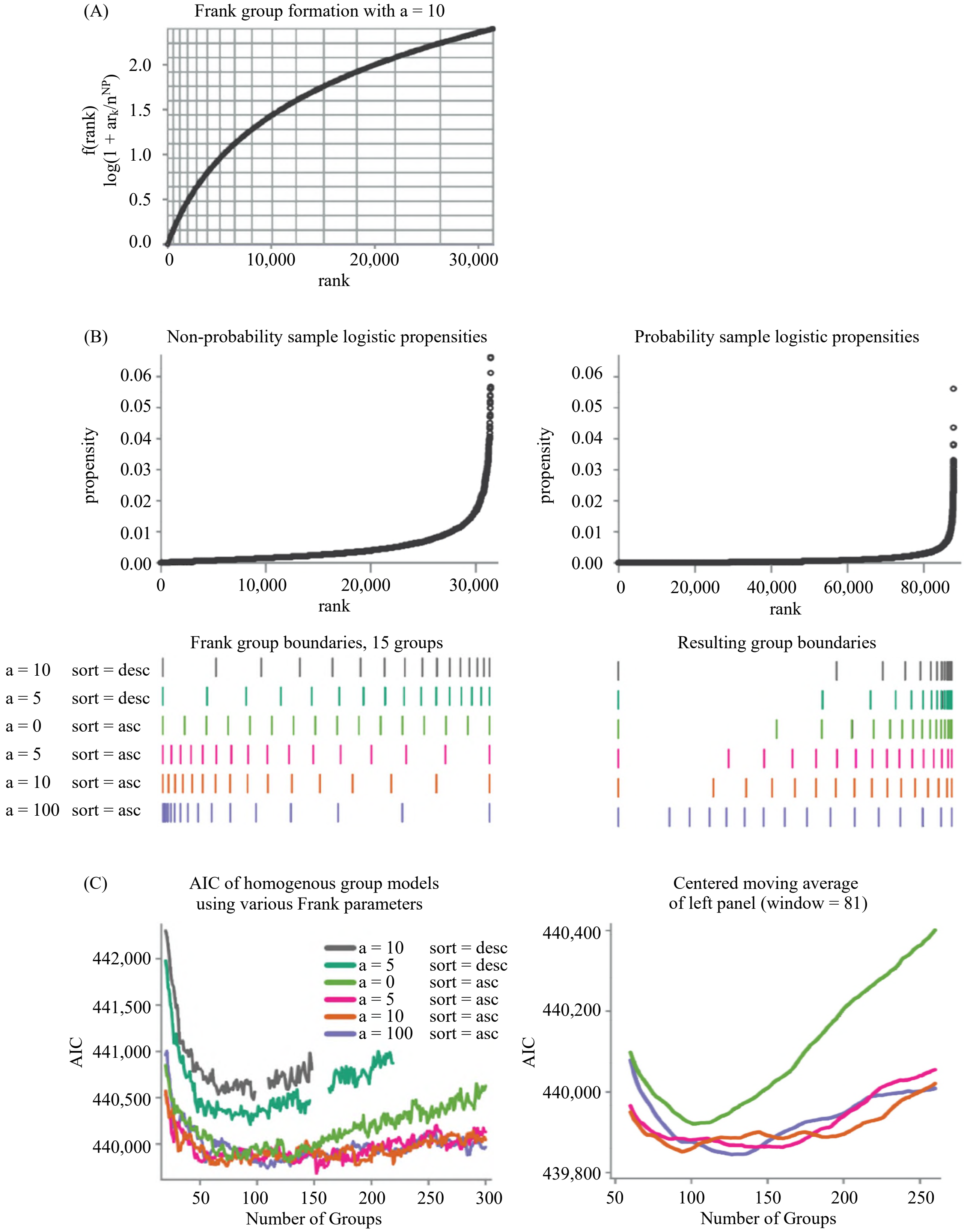
Figure A.1 below contains three sub-figures, Figures A.1(A), A.1(B) and A.1(C), that illustrate the behaviour of the Frank method for the data described in Section 6.1 and for method 5 described in Section 6.2 when the LFS is used as the probability sample. The description of each sub-figure is provided below:

(A) Frank method with $a = 10,$ $G = 15$ and n^{NP} = 31,415. The rank, $r_k,$ is on the horizontal axis and the function of the rank, $f(r_k) = \log(1 + a\, r_k / n^{\text{NP}}),$ is on the vertical axis. The bins are equal-width in the range of $f(r_k).$ The constant a determines the shape of the function. As a increases, it becomes increasingly non-linear and the groups are more bunched to one side.

(B) The top panels show the sorted values of $\hat{p}_k^{\text{logistic}}$ for the non-probability (left) and probability (right) samples. Fifteen groups are formed based on the non-probability sample using Frank with different values of a and both sorting orders, resulting in different group boundaries as represented by the coloured bars in the bottom panels. For the non-probability sample (bottom left panel), when the rank is defined in ascending order of $\hat{p}_k^{\text{logistic}},$ the groups are smaller for small values of $\hat{p}_k^{\text{logistic}}.$ When the rank is defined in descending order of $\hat{p}_k^{\text{logistic}},$ the groups are smaller for large values of $\hat{p}_k^{\text{logistic}}.$ Increasing a increases the bunching, while $a = 0$ gives equal-sized groups.

(C) The AIC (3.15) versus the number of groups for different values of a and both sorting orders. Sorting $\hat{p}_k^{\text{logistic}}$ in ascending order leads to smaller values of AIC, without much sensitivity to changes in the value of $a.$ The AIC is minimized with around 100 groups for all parameterizations. The right panel smooths the left panel using a centered moving average filter with window size 81. The smoothed curves show the Frank method performs slightly better than equal-sized groups $(a = 0),$ especially when the number of groups is higher than optimal, adding some robustness to the choice of the number of groups. When the number of groups is large and $\hat{p}_k^{\text{logistic}}$ are sorted in descending order, it occurs that some groups do not contain any probability sample unit. As a result, $\hat{p}_g$ is undefined for those groups, and the AIC cannot be computed.

Figure A.1 Illustration of the Frank method.



Appendix 3

Auxiliary variables

Age Group: 5-year age groups, starting from 15-19 and ending with 75+.

Sex: Male/Female.

Education: 8 categories (Less than high school; High school; Some post-secondary; Trades certificate or diploma; Community college, CEGEP, etc.; University certificate below Bachelor's; Bachelor's degree; Above Bachelor's degree).

Economic Region: Sub-provincial geography partitioning the country. It contains 73 levels, but some were collapsed due to insufficient respondent counts; 56 levels were used in the models.

Immigration: 3 levels (Born in Canada; Landed immigrant; Not a landed immigrant).

Household Size: Number of people in the household, regardless of age, capped at 6.

Marital Status: 6 levels (Married; Common-law; Widow or widower; Separated; Divorced; Single, Never married).

Employment Status: 3 levels (Employed and at work at least part of the reference week; Employed but absent from work; Not employed).

References

Bahamyirou, A., and Schnitzer, M.E. (2021). Data integration through outcome adaptive LASSO and a collaborative propensity score approach. *arXiv preprint arXiv:2103.15218*.

Baribeau, B. (2020). Trial by COVID for Statistics Canada's web panel pilot. Internal document, Statistics Canada.

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.

Beaumont, J.-F., and Émond, N. (2022). A bootstrap variance estimation method for multistage sampling and two-phase sampling when Poisson sampling is used at the second phase. *Stats*, 5, 339-357.

Beaumont, J.-F., and Patak, Z. (2012). On the generalized bootstrap for sample surveys with special attention to Poisson sampling. *International Statistical Review*, 80, 127-148.

Breiman, L., Friedman, J.H., Olshen, R.A. and Stone, C.J. (1984). *Classification and Regression Trees*. Wadsworth, Boca Raton, FL.

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

Chu, K., and Beaumont, J.-F. (2019). The use of classification trees to reduce selection bias for a non-probability sample with help from a probability sample. *Proceedings of the Survey Methods Section*, Statistical Society of Canada, May 2019.

Deville, J.-C., and Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87, 376-382.

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

Eltinge, J.L., and Yansaneh, I.S. (1997). Diagnostics for formation of nonresponse adjustment cells, with an application to income nonresponse in the U.S. Consumer Expenditure Survey. *Survey Methodology*, 23, 1, 33-40. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/1997001/article/3103-eng.pdf.

Ferri-Garcia, R., Beaumont, J.-F., Bosa, K., Charlebois, J. and Chu, K. (2021). Weight smoothing for nonprobability surveys. *TEST* (published online).

Ferri-Garcia, R., and Rueda, M.d.M. (2018). Efficiency of propensity score adjustment and calibration on the estimation from non-probabilistic online surveys. *SORT*, 42, 159-182.

Gambino, J., Kennedy, B., and Singh, M.P. (2001). Regression composite estimation for the Canadian Labour Force Survey: Evaluation and implementation. *Survey Methodology*, 27, 1, 65-74. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2001001/article/5855-eng.pdf.

Haziza, D., and Lesage, É. (2016). A discussion of weighting procedures for unit nonresponse. *Journal of Official Statistics*, 32, 129-145.

Izenman, A.J. (2008). *Modern Multivariate Statistical Techniques: Regression, Classification, and Manifold Learning.* New York: Springer Science & Business Media.

Lee, S. (2006). Propensity score adjustment as a weighting scheme for volunteer panel Web survey. *Journal of Official Statistics*, 22, 329-349.

Lohr, S.L. (2021). Multiple-frame surveys for a multiple-data-source world. *Survey Methodology*, 47, 2, 229-263. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2021002/article/00008-eng.pdf.

Lohr, S., Hsu, V. and Montaquila, J. (2015). Using classification and regression trees to model survey nonresponse. *Proceedings of the Survey Research Methods Section*, American Statistical Association, Alexandria, VA.

Lumley, T., and Scott, A. (2015). AIC and BIC for modeling with complex survey data. *Journal of Survey Statistics and Methodology*, 3, 1-18.

Phipps, P., and Toth, D. (2012). Analyzing establishment nonresponse using an interpretable regression tree model with linked administrative data. *The Annals of Applied Statistics*, 6, 772-794.

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

Rao, J.N.K., Wu, C.F.J. and Yue, K. (1992). Some recent work on resampling methods for complex surveys. *Survey Methodology*, 18, 2, 209-217. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/1992002/article/14486-eng.pdf.

Renaud, M., and Beaumont, J.-F. (2020). Crowdsourcing during a pandemic: The Statistics Canada experience. Paper presented at the Advisory Committee on Statistical Methods, Statistics Canada, October 27, 2020.

Rivers, D. (2007). Sampling from web surveys. *Proceedings of the Survey Research Methods Section*, American Statistical Association, Alexandria, VA.

Royall, R.M. (1970). On finite population sampling theory under certain linear regression models. *Biometrika*, 57, 377-387.

Valliant, R. (2020). Comparing alternatives for estimation from nonprobability samples. *Journal of Survey Statistics and Methodology*, 8, 2, 231-263.

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

Valliant, R., Dorfman, A.H. and Royall, R.M. (2000). *Finite Population Sampling and Inference: A Prediction Approach*. New-York: John Wiley & Sons, Inc.

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.

Wu, C. (2022). Statistical inference with non-probability survey samples. *Survey Methodology*, 48, 2, 283-311. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2022002/article/00002-eng.pdf.

Yang, S., and Kim, J.K. (2020). Statistical data integration in survey sampling: A review. *Japanese Journal of Statistics and Data Science*, 3, 625-650.

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, 1, 29-58. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2021001/article/00004-eng.pdf.

Comments on "Handling non-probability samples through inverse probability weighting with an application to Statistics Canada's crowdsourcing data"

Julie Gershunskaya and Vladislav Beresovsky[1]

Abstract

Beaumont, Bosa, Brennan, Charlebois and Chu (2024) propose innovative model selection approaches for estimation of participation probabilities for non-probability sample units. We focus our discussion on the choice of a likelihood and parameterization of the model, which are key for the effectiveness of the techniques developed in the paper. We consider alternative likelihood and pseudo-likelihood based methods for estimation of participation probabilities and present simulations implementing and comparing the AIC based variable selection. We demonstrate that, under important practical scenarios, the approach based on a likelihood formulated over the observed pooled non-probability and probability samples performed better than the pseudo-likelihood based alternatives. The contrast in sensitivity of the AIC criteria is especially large for small probability sample sizes and low overlap in covariates domains.

Key Words: Non-probability sample; Participation probabilities; Sample likelihood; Data combining.

Building on recent developments in data integration methods, Beaumont et al. (2024) propose and apply to real data innovative model selection approaches for estimation of participation probabilities for non-probability sample units. We congratulate the authors for their inspiring and timely contribution and appreciate the opportunity to comment and discuss the methods considered in the paper.

Survey statisticians are increasingly faced with the need to extract useful information from data collected without a well-planned probability survey design. At the same time, we witness rapid developments of machine learning methods capable to efficiently handle multidimensional sets of covariates. There is the growing realization that machine learning can be useful for handling estimation from non-probability samples. The current paper leads the way in adapting these methods for the combined non-probability and probability samples setup. The authors propose a general modified AIC formula that accounts for the probability sampling design in the combined samples setup. They also derive the AIC expression for the special case of homogeneous groups and apply it to choose among partitions in rank-based methods for forming the groups. Finally, the authors adapt the CART tree-growing algorithm by using a pseudo-likelihood as an objective function and applying the modified AIC to prune the tree.

We focus our discussion on the choice of a likelihood and parameterization of the model, which are key for the effectiveness of the techniques developed in the paper. In Section 1, we review several approaches for estimation of participation probabilities proposed in recent years and provide AIC expressions for the homogeneous groups case. In Section 2, we present simulations implementing and comparing the AIC based variable selection for these approaches. We provide some concluding remarks in Section 3.

[1] Julie Gershunskaya and Vladislav Beresovsky, U.S. Bureau of Labor Statistics, 2 Massachusetts Ave NE Washington, DC 20212, U.S.A. E-mail: Gershunskaya.Julie@bls.gov.

1. Approaches to estimation of participation probabilities

We start with the two pseudo-likelihood based approaches considered in the paper, then we discuss an exact likelihood based approach and propose a modification to the ALP method of Wang, Valliant and Li (2021), see also a related discussion in Gershunskaya and Lahiri (2023). We then consider the case of homogeneous groups and derive and compare the AIC expressions for each of the approaches for this important and relatively simple special case. Throughout, unless explicitly stated, we follow the notation of the current paper.

1.1 CLW method

Assuming that both non-probability and probability samples are selected from the same finite population U, Chen, Li and Wu (2020) (hereafter CLW) start by writing a log-likelihood over units in U, with respect to Bernoulli variable δ_k:

$$\ell^{\text{CLW}}(\boldsymbol{\alpha})=\sum_{k\in U}\left\{\delta_k\log\left[p_k(\boldsymbol{\alpha})\right]+(1-\delta_k)\log\left[1-p_k(\boldsymbol{\alpha})\right]\right\},\tag{1.1}$$

where δ_k is unit's k non-probability sample inclusion indicator, $p_k(\mathbf{x_k})=P\{\delta_k=1\,|\,\mathbf{x_k}\}$, and $\boldsymbol{\alpha}$ is the parameter vector in a logistic regression model for p_k, where $\text{logit}(p_k(\boldsymbol{\alpha}))=\boldsymbol{\alpha}^T\mathbf{x_k}$.

Since finite population units are not observed, CLW re-group the sum in (1.1) by presenting it as a sum of two parts: part 1 involves the sum over the non-probability sample units, s_{NP}, and part 2 is the sum over finite population U:

$$\ell^{\text{CLW}}(\boldsymbol{\alpha})=\sum_{k\in s_{\text{NP}}}\log\left[\frac{p_k(\boldsymbol{\alpha})}{1-p_k(\boldsymbol{\alpha})}\right]+\sum_{k\in U}\log\left[1-p_k(\boldsymbol{\alpha})\right].\tag{1.2}$$

CLW employ a pseudo-likelihood approach by replacing the sum over the finite population with its probability sample based estimate:

$$\hat{\ell}^{\text{CLW}}(\boldsymbol{\alpha})=\sum_{k\in s_{\text{NP}}}\log\left[\frac{p_k(\boldsymbol{\alpha})}{1-p_k(\boldsymbol{\alpha})}\right]+\sum_{k\in s_P}w_k\log\left[1-p_k(\boldsymbol{\alpha})\right],\tag{1.3}$$

where weights $w_k=\pi_k^{-1}$ are inverse values of the reference sample inclusion probabilities π_k. Estimates are obtained by solving respective pseudo-likelihood based estimating equations.

Note that the likelihood under this approach is formulated with respect to indicator δ_k, although this variable is not observed.

1.2 ALP method

For their Adjusted Logistic Propensity (ALP) weighting method, Wang et al. (2021) introduce an imaginary construct consisting of two parts: they *stack* together non-probability sample s_{NP} (part 1) and finite population U (part 2). Since non-probability sample units belong to the finite population, they appear in the stacked set twice. They formulate a Bernoulli likelihood for variable R_k, where $R_k = 1$ if unit k belongs to part 1 of the stacked set; and $R_k = 0$ otherwise:

$$\ell^{\mathrm{ALP}}(\gamma) = \sum_{k \in s_{\mathrm{NP}}} \log\left[p_{Rk}(\gamma)\right] + \sum_{k \in U} \log\left[1 - p_{Rk}(\gamma)\right], \tag{1.4}$$

where γ is the parameter vector in a logistic regression model for $p_{Rk}(\mathbf{x_k}) = P\{R_k = 1 \mid \mathbf{x_k}\}$. Since the finite population is not available, they apply a pseudo-likelihood approach:

$$\hat{\ell}^{\mathrm{ALP}}(\gamma) = \sum_{k \in s_{\mathrm{NP}}} \log\left[p_{Rk}(\gamma)\right] + \sum_{k \in s_P} w_k \log\left[1 - p_{Rk}(\gamma)\right], \tag{1.5}$$

leading to an estimate of p_{Rk}. However, the actual goal is to find probabilities p_k rather than p_{Rk}. At the second step of the ALP method, estimates of p_k are derived from identity

$$p_{Rk} = \frac{p_k}{1 + p_k}. \tag{1.6}$$

Wang et al. (2021) noted that in their simulations the ALP estimator was more efficient than the CLW, especially when the non-probability sample size was much larger than the probability sample size.

1.3 ILR method

Let us now consider an exact likelihood approach formulated over the pooled non-probability and probability samples. Savitsky, Williams, Gershunskaya and Beresovsky (2023) propose to stack together the two samples and consider indicator variable $z_k = 1$ if unit k belongs to the non-probability sample (part 1), and $z_k = 0$ if unit k belongs to the probability sample (part 2). Under this stacked sample construction, if there is an overlap between the two samples, s_{NP} and s_P, then the overlapping units are included into the stacked set, s, twice: once as part of the non-probability sample (with $z_k = 1$) and once as part of the reference probability sample (with $z_k = 0$). We do not need to know which units overlap or whether there are any overlapping units. They use first principles to prove a relationship between probabilities $p_{zk}(\mathbf{x_k}) = P\{z_k = 1 \mid \mathbf{x_k}\}$ of being in part 1 of the stacked set, on the one hand, and inclusion probabilities, p_k and π_k, on the other hand:

$$p_{zk} = \frac{p_k}{\pi_k + p_k}. \tag{1.7}$$

Elliott (2009) and Elliott and Valliant (2017) derived expression (1.7) under the assumption of non-overlapping samples s_{NP} and s_P. The derivation given in Savitsky et al. (2023) does not require this assumption.

To obtain estimates of p_k, Beresovsky (2019) proposed an approach, labeled Implicit Logistic Regression (ILR), to allow the estimation of p_k directly from the likelihood formulated on the combined sample. In ILR, probabilities $p_k = p_k(\alpha)$ are parameterized as $\operatorname{logit}(p_k(\alpha)) = \alpha^T \mathbf{x_k}$ (as in CLW), and identity (1.7) is used to present p_{zk} as a composite function of $\boldsymbol{\alpha}$; that is, $p_{zk} = p_{zk}(p_k(\boldsymbol{\alpha})) = p_k(\boldsymbol{\alpha})/(\pi_k + p_k(\boldsymbol{\alpha}))$. The log-likelihood for observed Bernoulli variable z_k is

$$\ell^{\text{ILR}}(\boldsymbol{\alpha}) = \sum_{k \in s_{\text{NP}}} \log\left[p_{zk}(p_k(\boldsymbol{\alpha}))\right] + \sum_{k \in s_P} \log\left[1 - p_{zk}\left(p_k(\boldsymbol{\alpha})\right)\right]. \tag{1.8}$$

The score equations are obtained from (1.8) by taking the derivatives, with respect to $\boldsymbol{\alpha}$, of composite function $p_{zk} = p_{zk}(p_k(\boldsymbol{\alpha}))$. This way, the estimates of p_k are obtained directly from (1.8) in a single step.

Note that for the ILR approach, the probability sample inclusion probabilities π_k are supposed to be known for all units in the combined set. This is possible for many probability surveys. If not immediately available, probabilities π_k for units in s_{NP} can be determined if probability sample design variables are available for non-probability sample units. As discussed in Elliott and Valliant (2017), π_k can be estimated using a regression model. Savitsky et al. (2023) used the Bayesian modeling technique to obtain both π_k and p_k. On the other hand, if probabilities π_k are not available for the non-probability part of the combined sample, one can apply a pseudo-likelihood approach labeled “pseudo-ILR”, as discussed below in Section 1.4.

1.4 Pseudo-ILR method

The estimation method of Wang et al. (2021) can be modified to a one-step estimation procedure similar to ILR: p_k can be parameterized using the logistic link function as $\text{logit}(p_k(\boldsymbol{\alpha})) = \boldsymbol{\alpha}^T \mathbf{x_k}$, while probabilities p_{Rk} in (1.6) could be viewed as a composite function, $p_{Rk} = p_{Rk}(p_k(\boldsymbol{\alpha})) = p_k(\boldsymbol{\alpha})/(1 + p_k(\boldsymbol{\alpha}))$. The pseudo-likelihood takes the form:

$$\hat{\ell}^{\text{PILR}}(\boldsymbol{\alpha}) = \sum_{k \in s_{\text{NP}}} \log\left[p_{Rk}(p_k(\boldsymbol{\alpha}))\right] + \sum_{k \in s_P} w_k \log\left[1 - p_{Rk}(p_k(\boldsymbol{\alpha}))\right]. \tag{1.9}$$

This change in estimation of model parameters makes the approach more efficient and less biased than ALP. It also avoids cases where estimates of p_k become greater than 1, as may occur under the ALP approach where the estimation is performed in two steps.

Note that pseudo-likelihood based (1.3) and (1.9) use exactly the same set of observed data and yet these expressions are quite different. We expect the pseudo-ILR approach to give more efficient estimates because it is based on a likelihood properly formulated with respect to observed Bernoulli variable R_k, while the CLW likelihood is given with respect to unobserved δ_k. Our simulations (not included in the discussion) confirm a better performance of the pseudo-ILR relative to the CLW approach. The effect on the AIC performance is shown in simulations Section 2.

1.5 Homogeneous groups

The authors presented the log-likelihood and AIC expressions under the CLW approach for the special case of homogeneous groups. We now extend their approach to pseudo-ILR and ILR methods.

For the pseudo-ILR approach, it is easy to see that, for a given partition, estimates of group g participation probabilities p_g are $\hat{p}_g = n_g^{\mathrm{NP}} / \hat{N}_g$, where $\hat{N}_g = \sum_{k \in s_P} w_k$ (the same as in the CLW approach.) The AIC is given by

$$\mathrm{AIC}^{\mathrm{PILR}} = -2\hat{\ell}^{\mathrm{PILR}}(\hat{\alpha}) + 2G + 2\sum_{g=1}^{G} \left[\frac{1-\hat{p}_g}{1+\hat{p}_g}\right] n_g^{\mathrm{NP}} \frac{\left[\hat{cv}_d(\hat{N}_g)\right]^2}{1-\hat{p}_g}, \tag{1.10}$$

where the log-likelihood is

$$\hat{\ell}^{\mathrm{PILR}}(\hat{\alpha}) = \sum_{g=1}^{G} \hat{N}_g \left[\hat{p}_g \log \frac{\hat{p}_g}{1+\hat{p}_g} + \log \frac{1}{1+\hat{p}_g}\right]. \tag{1.11}$$

Note that last terms in formulas for AIC under the CLW and pseudo-ILR approaches differ by a factor $(1-\hat{p}_g)/(1+\hat{p}_g) < 1.$ That is, for a given partition, the penalty term in the pseudo-ILR approach is always smaller than the one in the CLW approach.

In the ILR approach, estimates of p_g are not available in closed form. They can be found by solving equations:

$$\sum_{k \in s_g} \frac{\pi_{gk}}{\pi_{gk} + p_g} = n_g^P, \tag{1.12}$$

where π_{gk} are assumed known, s_g is the part of combined sample that belongs to group g, $g = 1, \ldots, G.$ Since ILR is based on an exact likelihood, the AIC formula for ILR does not have the third term and is a standard AIC expression:

$$\mathrm{AIC}^{\mathrm{ILR}} = -2\hat{\ell}^{\mathrm{ILR}}(\hat{\alpha}) + 2G, \tag{1.13}$$

where

$$\hat{\ell}^{\mathrm{ILR}}(\hat{\alpha}) = \sum_{g=1}^{G} \left[\sum_{k \in s_g^{\mathrm{NP}}} \log \frac{\hat{p}_g}{\pi_{gk} + \hat{p}_g} + \sum_{k \in s_g^P} \log \frac{\pi_{gk}}{\pi_{gk} + \hat{p}_g}\right]. \tag{1.14}$$

Let us compare the penalty terms of the three methods for a given set of homogeneous groups. (The partitioning itself depends on the likelihood used, but we do not consider this effect at the moment.) We suppose that, all other factors being equal, the smaller the penalty term, the better AIC works. Thus, we expect the ILR based criteria to perform better than the pseudo-ILR or CLW. In turn, the pseudo-ILR is expected to work better than CLW, especially when the non-probability sample is relatively large and $\hat{p}_g$ in (1.10) and in formula (3.15) of the discussed paper are closer to 1. Simulations results of Section 2 support this reasoning.

2. Simulations

We conducted a simulation experiment to study performances of the AIC based on CLW, pseudo-ILR, and ILR approaches.

For each unit $k = 1\ldots, N$ in a finite population U of size $N =$ 10,000, we generated covariates x_{1k} and x_{2k} as independent standard normal variables.

We use Poisson sampling with participation probabilities p_k to select non-probability sample from population U. Probabilities p_k are generated as

$$\operatorname{logit}(p_k) = \alpha_0 + \alpha_1 x_{1k} + \alpha_2 x_{2k}, \tag{2.1}$$

where we set specific coefficient values for different simulation scenarios, as follows:

- setting $\alpha_0 = -5$ produces sample s_{NP} of approximate size $n^{\text{NP}} = 100$; setting $\alpha_0 = -2.5$ produces sample s_{NP} of approximate size $n^{\text{NP}} =$ 1,000;
- α_1 was set to 1 for all scenarios;
- α_2 is set to values on a grid from $-$ 0.3 to 0.3, corresponding to a series of scenarios. Note that setting $\alpha_2 = 0$ corresponds to the case where p_k is independent of x_{2k}; larger values of α_2 produce stronger dependence on x_{2k}.

For probability sample s_P, we considered scenarios where sample sizes are $n^P =$ 100 or $n^P =$ 1,000. The probability sample is generated using the probability proportional to size (PPS) without replacement design, where the measure of size is defined as

$$\operatorname{logit}(m_k) = 1 + \beta x_{1k}. \tag{2.2}$$

In multivariate models with a large number of auxiliary variables and interactions, it is likely that non-probability and probability samples would have very little overlap in some of the variable-defined domains. Firth (1993) and Heinze and Schemper (2002) demonstrated that little overlap, or separation, may result in unstable estimates of model parameters. This is the case where it is essential to use an effective method for variables selection. Therefore, in our simulations, we included scenarios of low and high overlap in variables domains. Values of coefficient β are set to regulate the degree of the overlap across the range of covariate x_1. To simulate the "high" overlap, we set $\beta = 1$ (so that $\beta = \alpha_1$); for the "low" overlap scenario, we set $\beta = -1$.

Table 2.1 presents a summary of considered simulation scenarios, S1-S4, characterized by combinations of high or low overlap and different sample sizes. We applied three approaches, CLW, pseudo-ILR, and ILR, for each scenario to choose between two models:

$$\begin{gathered}\text{Full Model: } \operatorname{logit}(p_k) = \alpha_0 + \alpha_1 x_{1k} + \alpha_2 x_{2k}, \\ \text{Abbreviated Model: } \operatorname{logit}(p_k) = \alpha_0 + \alpha_1 x_{1k}.\end{gathered} \tag{2.3}$$

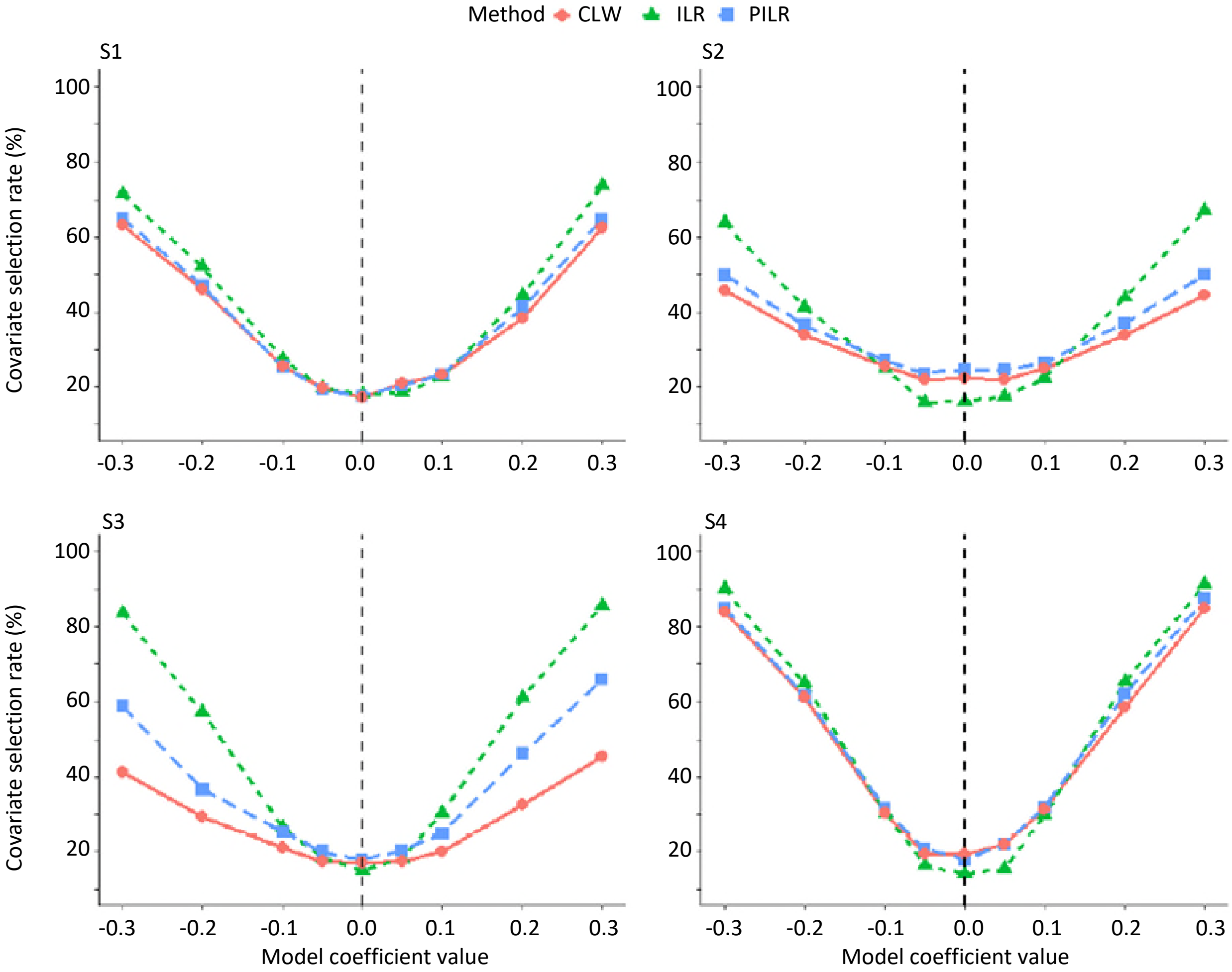
Table 2.1
Summary of considered simulation scenarios.

	Overlap	n^P	n^{NP}
S1	High	100	100
S2	Low	100	100
S3	Low	100	1,000
S4	Low	1,000	100

In each case, we compute AIC for the Full and Abbreviated models and choose the model with the lower AIC. We repeat this test $B = 500$ times for each scenario and the value of α_2 and find the percentage of times r the Full Model is chosen, $p = 100r/B$.

Plots in Figure 2.1 correspond to the four scenarios in Table 2.1. We plot percentage p against the values of coefficient α_2. For larger absolute values of α_2, the higher percentage p would be preferable; when α_2 is close to 0, the lower values of p would indicate a better performance. The line with red dots shows the CLW results, the line with blue squares is for pseudo-ILR, and the line with green triangles shows results for ILR.

Figure 2.1 Relative AIC performances under S1-S4 scenarios for CLW (red dots), ILR (green triangles), and pseudo-ILR (blue squares).



For all scenarios considered, the ILR approach performs the best: for larger absolute values of α_2, the ILR based AIC most frequently chooses the Full Model; when α_2 is closer to 0, the ILR based criteria selects the Full Model the least number of times. In the high overlap (S1) or relatively large probability sample size (S4) scenarios, all three approaches produce close results. However, when the probability sample is relatively small and the overlap is low (S2 and S3), the performance of ILR based test is significantly better than the other two methods. In most cases, the pseudo-ILR based test is slightly better than the CLW approach. When the non-probability sample size is large relative to the probability sample size (S3), the difference in performances increases: when absolute value of α_2 is close to 0.3, the CLW based criteria chooses the Full Model in only about 40-50% of times, whereas the pseudo-ILR based criteria chooses it in about 60%, and the ILR based test chooses it in roughly 85% of cases.

3. Concluding remarks

We commend the authors on their contribution in adapting model selection algorithms to data integration problems. The choice of an objective function is important for this task. In our discussion, we considered several recently proposed alternative likelihood functions. The exact likelihood based ILR method performed better than the pseudo-likelihood based alternatives in the important practical situation of small probability sample sizes and low overlap in covariates domains.

We note disadvantage of the CLW method when the probability sample is small and the non-probability sample is relatively large. In this case, we also noticed convergence problems with the Newton-Raphson algorithm in the CLW method.

The ILR requires that the probability sample inclusion probabilities be available for non-probability sample units. If these probabilities can be derived based on available design variables, then ILR would be the preferred method. Otherwise, the pseudo-ILR appears to be a viable option.

References

Beaumont, J.-F., Bosa, K., Brennan, A., Charlebois, J. and Chu, K. (2024). Handling non-probability samples through inverse probability weighting with an application to Statistics Canada's crowdsourcing data. *Survey Methodology*, 50, 1, 77-106. Paper available at http://www.statcan.gc.ca/pub/12-001-x/2024001/article/00004-eng.pdf.

Beresovsky, V. (2019). On application of a response propensity model to estimation from web samples. In ResearchGate.

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.R. (2009). Combining data from probability and non-probability samples using pseudo-weights. *Survey Practice*, 2, 813-845.

Elliott, M.R., and Valliant, R. (2017). Inference for nonprobability samples. *Statistical Science*, 32(2), 249-264.

Firth, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika*, 80(1), 27-38.

Gershunskaya, J., and Lahiri, P. (2023). Discussion of “Probability vs. nonprobability sampling: From the birth of survey sampling to the present day”, by Graham Kalton. *Statistics in Transition New Series*, 24(3).

Heinze, G., and Schemper, M. (2002). A solution to the problem of separation in logistic regression. *Statistics in Medicine*, 21, 2409-2419.

Savitsky, T.D., Williams, M.R., Gershunskaya, J. and Beresovsky, V. (2023). Methods for combining probability and nonprobability samples under unknown overlaps. *Statistics in Transition New Series*, 24(5), 1-34.

Wang, L., Valliant, R. and Li, Y. (2021). Adjusted logistic propensity weighting methods for population inference using nonprobability volunteer-based epidemiologic cohorts. *Stat Med.*, 40(4), 5237-5250.

Comments on "Handling non-probability samples through inverse probability weighting with an application to Statistics Canada's crowdsourcing data"

Changbao Wu[1]

Abstract

We provide comparisons among three parametric methods for the estimation of participation probabilities and some brief comments on homogeneous groups and post-stratification.

Key Words: Inverse probability weighting; Participation probability; Pooled sample; Poststratification; Pseudo maximum likelihood; Reference probability sample.

Beaumont, Bosa, Brennan, Charlebois and Chu (2024) provided a thorough examination of inverse probability weighting methods for non-probability samples, including parametric approaches and tree-based classification methods, with a major focus on variable selection. This is an important research topic, as the demands on using non-probability samples in applied fields and official statistics have been steadily increasing in recent years. The current paper is a timely contribution to investigating and comparing different methodologies using a real world dataset. I would like to thank the Guest Editor, Dr. Partha Lahiri, for the invitation and I appreciate the opportunity for a short discussion. In Section 1, I provide comparisons among three parametric methods for the estimation of participation probabilities and for inverse probability weighting. I also provide some brief comments on the use of homogeneous groups for post-stratification in Section 2.

1. The methods of Chen, Li and Wu (2020), Valliant and Dever (2011), and Wang, Valliant and Li (2021)

These are three parametric methods frequently cited in recent literature on inverse probability weighting (IPW) using estimated participation probabilities for non-probability samples. There are conceptual differences among the three methods as well as similarities in numeric results when sample sizes are small relative to the population size.

The foundation for IPW estimators has been built under probability sampling with the Horvitz-Thompson estimator. Follow the notation of Beaumont et al. (2024), let $U = \{1, 2, \ldots, N\}$ be the finite population of size N. Let d denote the probability sampling design for selecting a probability sample s. Under the probability sampling design, the sample inclusion indicator variable $\delta_k = I(k \in s)$ is defined for every unit k in the target population U, i.e., for $k = 1, 2, \ldots, N$, where $I(\cdot)$ is the indicator function, and the sampling inclusion probabilities $\pi_k = P(\delta_k = 1 \mid U) = P(k \in s \mid U)$ can be computed based on the given

[1] Changbao Wu, Department of Statistics and Actuarial Science, University of Waterloo, Waterloo ON N2L 3G1. E-mail: cbwu@uwaterloo.ca.

sampling design, d. The joint distribution of $(\delta_1, \delta_2, \ldots, \delta_N)$ under repeated sampling characterizes the sampling design features, and the Horvitz-Thompson estimator $\hat{\theta}_{\text{HT}} = \sum_{k \in s} y_k / \pi_k$ for the population total $\theta = \sum_{k \in U} y_k$ is uniquely design-unbiased among a class of linear estimators. The establishment of this fundamental result in probability sampling involves (i) the positivity assumption, i.e., $\pi_k > 0$ for all k in U, so that $\hat{\theta}_{\text{HT}}$ can be re-written as $\hat{\theta}_{\text{HT}} = \sum_{k \in U} \delta_k y_k / \pi_k$; and (ii) the sample inclusion indicator δ_k is independent of y_k and $E_d(\delta_k \mid U) = \pi_k$, where the expectation E_d is with respect to the sampling design d.

Let s_{NP} be a non-probability sample of size n_{NP} from the population U. Let $\{(y_k, \mathbf{x}_k), k \in s_{\text{NP}}\}$ be the non-probability sample dataset. Once again, the sample participation indicator $\delta_k = I(k \in s_{\text{NP}})$ is defined for every unit k in the target population U, i.e., for $k = 1, 2, \ldots, N$. Unlike probability sampling, the participation probabilities $p_k = P(\delta_k = 1 \mid U)$ for the non-probability sample s_{NP} are unknown and hence need to be estimated, which requires assumptions on the form of p_k and an assumed model, denoted as q, for the participation mechanism. The model q leads to specifications of the joint distribution of $(\delta_1, \delta_2, \ldots, \delta_N)$. There are two commonly assumed components for model q: (i) the participation probabilities satisfy $p_k = P(\delta_k = 1 \mid \mathbf{x}_k, y_k) = P(\delta_k \mid \mathbf{x}_k) > 0$, $i = 1, 2, \ldots, N$; and (ii) the sample inculsion indicators $\delta_1, \delta_2, \ldots, \delta_N$ are conditionally independent given $(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N)$.

It should be emphasized that IPW estimators for non-probability samples need to be constructed and evaluated under the assumed model q on participation mechanism and under the induced distribution on $(\delta_1, \delta_2, \ldots, \delta_N)$. It is where conceptual differences among the three methods can clearly be identified. The method of Chen et al. (2020) for estimating p_k is based on the full likelihood function $\prod_{k=1}^{N} p_k^{\delta_k} (1 - p_k)^{1-\delta_k}$. With a pre-specified parametric form $p_k = p(\mathbf{x}_k, \boldsymbol{\alpha})$, the maximum pseudo likelihood estimator $\hat{\boldsymbol{\alpha}}$ is derived and assessed under an assumed parametric model q on $(\delta_1, \delta_2, \ldots, \delta_N)$ as well as the sampling design, d, for the reference probability sample s_P. The IPW estimator $\hat{\theta}_{\text{IPW}} = \sum_{k \in s_{\text{NP}}} y_k / \hat{p}_k$, where $\hat{p}_k = p(\mathbf{x}_k, \hat{\boldsymbol{\alpha}})$, is consistent for θ under the joint randomization of model q and the sampling design d. The results are rigorously established with no restrictions on the "sampling fraction" n_{NP}/N or the parametric form for $p_k = p(\mathbf{x}_k, \boldsymbol{\alpha})$, whether it follows from a logistic regression model or any other models suitable for a binary response variable.

The paper by Valliant and Dever (2011) was the first serious attempt in addressing estimation of participation probabilities under the two-sample setup as described in Beaumont et al. (2024). It inspired several followup papers, including Chen et al. (2020) and Wang et al. (2021). The proposed method was based on fitting a survey weighted logistic regression model to the pooled sample $s_{\text{NP}} \cup s_P$ with the "response variable" defined as $D_k = 1$ if $k \in s_{\text{NP}}$ and $D_k = 0$ if $k \in s_P$, for $k \in s_{\text{NP}} \cup s_P$, assuming there are no overlaps between s_{NP} and s_P. It is apparent that the D_i's are defined with the given s_{NP} and s_P and are conceptually different from the sample participation indicators $(\delta_1, \delta_2, \ldots, \delta_N)$. Under an assumed parametric model ξ on the D_k's with $\pi(\mathbf{x}_k, \boldsymbol{\gamma}) = P(D_k = 1 \mid s_{\text{NP}} \cup s_P)$, the "theoretical IPW estimator" $\hat{\theta} = \sum_{k \in s_{\text{NP}}} y_k / \pi(\mathbf{x}_k, \boldsymbol{\gamma})$ should be evaluated first against model ξ, leading to $E_\xi(\hat{\theta} \mid s_{\text{NP}}, s_P) = E_\xi \left\{ \sum_{k \in s_{\text{NP}} \cup s_P} D_k y_k / \pi(\mathbf{x}_k, \boldsymbol{\gamma}) \mid s_{\text{NP}}, s_P \right\} = \sum_{k \in s_{\text{NP}} \cup s_P} y_k$. In creating a set of weights for the pooled sample

$s_{\text{NP}} \cup s_P$ for "survey-weighted" logistic regression analysis on the D_k's and without any prior knowledge of how s_{NP} was selected, Valliant and Dever (2011) simply assigned "1" to each $k \in s_{\text{NP}}$, which essentially assumes that units in s_{NP} are exchangeable with respect to model q. The IPW estimator of Valliant and Dever (2011), when assessed under model q for $(\delta_1, \delta_2, \ldots, \delta_N)$, is inconsistent unless s_{NP} is a simple random sample from the population as shown by Chen et al. (2020).

The more recent paper by Wang et al. (2021) adapted a strategy which is on the opposite direction of Valliant and Dever (2011). Instead of pooling the two samples together, the authors first created an enlarged population $s_{\text{NP}}^* \cup U$, where s_{NP}^* consists of the same set of units in $s_{\text{NP}} \subset U$ but these units are viewed differently in the union of s_{NP}^* and U. The authors defined the indicator variable $R_k = 1$ if $k \in s_{\text{NP}}^*$ and $R_k = 0$ if $k \in U$, and further defined the probability $\pi_k = P(R_k = 1 \mid s_{\text{NP}}^* \cup U) = P(k \in s_{\text{NP}}^* \mid k \in s_{\text{NP}}^* \cup U)$, for all $k \in s_{\text{NP}}^* \cup U$. I had some real difficulty in putting this formulation into a suitable conceptual framework, since the indicator variable R_k depends on $s_{\text{NP}} = \{i \mid i \in U \text{ and } \delta_i = 1\}$, which depends on the full vector of sample inclusion indicators $(\delta_1, \delta_2, \ldots, \delta_N)$. It is unclear what kind of probability model is behind $P(\cdot)$ in defining $\pi_k = P(R_k = 1 \mid s_{\text{NP}}^* \cup U)$. This further led to my struggle in understanding the arguments behind the identity $p_k = P(\delta_k = 1 \mid U) = \pi_k / (1 - \pi_k)$ (Wang et al., 2021, page 5241, equation (9)). The identity implies that a logistic regression model on the R_k's would lead to a model on the δ_k's with the log-link function, a potential source of concerns when the sampling fraction n_{NP}/N is large.

It turns out that the three methods produce similar numeric results for the estimated participation probabilities when the sampling fraction n_{NP}/N is small. This can be explained by checking computational details for each of the methods. Under a parametric model $p_k = p(\mathbf{x}_k, \boldsymbol{\alpha})$ and the assumption of conditional independence of $(\delta_1, \delta_2, \ldots, \delta_N)$ given $(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N)$, the full log-likelihood function for α is given by

$$\ell(\boldsymbol{\alpha}) = \sum_{k \in s_{\text{NP}}} \log\{p(\mathbf{x}_k, \boldsymbol{\alpha})\} + \sum_{k \in U - s_{\text{NP}}} \log\{1 - p(\mathbf{x}_k, \boldsymbol{\alpha})\}. \tag{1.1}$$

The second term on the right hand side of (1.1), denoted as

$$L_2 = \sum_{k \in U - s_{\text{NP}}} \log\{1 - p(\mathbf{x}_k, \boldsymbol{\alpha})\},$$

is not computable based on the non-probability sample s_{NP} since it involves $\mathbf{x}_k$ for all k not in the sample s_{NP}. The three methods, although conceptually distinctive, differ computationally only in terms of how the term L_2 is handled.

Let $\{(\mathbf{x}_k, d_k), k \in s_P\}$ be the reference probability sample dataset, where the d_k's are the survey weights for s_P. Treating L_2 as a total of a population of size $N - n_{\text{NP}}$, the method of Valliant and Dever (2011) is equivalent to estimating L_2 by

$$\hat{L}_2^{(1)} = \sum_{k \in s_P} w_k \log\{1 - p(\mathbf{x}_k, \boldsymbol{\alpha})\},$$

where $w_k = d_k(\hat{N} - n_{\text{NP}})/\hat{N}$ are the rescaled weights satisfying $\sum_{k \in s_P} w_k = \hat{N} - n_{\text{NP}}$ and $\hat{N} = \sum_{k \in s_P} d_k$ is the estimated population size for U. Noting that L_2 can be re-written as $L_2 = \sum_{k \in U} \log\{1 - p(\mathbf{x}_k, \boldsymbol{\alpha})\} - \sum_{k \in s_{\text{NP}}} \log\{1 - p(\mathbf{x}_k, \boldsymbol{\alpha})\}$, the method of Chen et al. (2020) replaces L_2 by

$$\hat{L}_2^{(2)} = \sum_{k \in s_P} d_k \log\{1 - p(\mathbf{x}_k, \boldsymbol{\alpha})\} - \sum_{k \in s_{\text{NP}}} \log\{1 - p(\mathbf{x}_k, \boldsymbol{\alpha})\},$$

which is design-unbiased (or design-consistent, depending on whether the d_k's are the basic design weights or calibrated/adjusted weights) regardless of the sampling fraction n_{NP}/N. The method of Wang et al. (2021) for estimating $\pi_k = P(R_k = 1 \mid s_{\text{NP}} \cup U)$ amounts to replacing L_2 by

$$\hat{L}_2^{(3)} = \sum_{k \in s_P} d_k \log\{1 - p(\mathbf{x}_k, \boldsymbol{\alpha})\}.$$

This clearly overshoots the target since $\hat{L}_2^{(3)}$ is an estimate for $\sum_{k \in U} \log\{1 - p(\mathbf{x}_k, \boldsymbol{\alpha})\}$ (or "undershoots" if we consider the fact that $\log\{1 - p(\mathbf{x}_k, \boldsymbol{\alpha})\} < 0$ for all k). However, the use of $\hat{L}_2^{(3)}$ to replace L_2 in $\ell(\boldsymbol{\alpha})$ results in a log-likelihood function that resembles a hypothetical scenario where the sample s_{NP} is taken from a larger population $s_{\text{NP}} \cup U$. The resulting π_k needs to be adjusted to make it closer to the actual participation probability p_k.

It is apparent that the three quantities $\hat{L}_2^{(1)}$, $\hat{L}_2^{(2)}$ and $\hat{L}_2^{(3)}$ do not differ too much when the sampling fraction n_{NP}/N is small, leading to similar estimated participation probabilities under these scenarios. The final adjustment step from the method of Wang et al. (2021), i.e., $p_k = \pi_k/(1 - \pi_k)$, also gives similar results, since we typically have $\pi_k = O(n_{\text{NP}}/N)$, which implies $p_k/\pi_k = 1 + o(1)$ when $n_{\text{NP}}/N = o(1)$.

2. Homogeneous groups and post-stratification

In practice, auxiliary variables which are included in probability or non-probability surveys are often categorical or ordinal, especially for surveys on human populations where basic information on demographic variables and social-economic indicators is routinely collected. When relevant variables for characterizing the participation mechanism are discrete, the IPW estimator is equivalent to a post-stratified estimator; see, for instance, Section 5 of Wu (2022) for a detailed discussion. A post-stratified (IPW) estimator uses uniform participation probabilities within the same post-stratum, which effectively removes the impact of extreme values of estimated participation probabilities often appearing with a parametric model when there are continuous auxiliary variables, and the estimator has a simple and easy-to-use form.

There are two major challenges, however, in forming homogeneous groups as post-strata. The first is the large number of initial groups when there are many discrete auxiliary variables that are available in the datasets. The variable selection methods discussed in Beaumont et al. (2024) have the potential to be practically useful in reducing the number of groups for the final IPW estimator. The second are scenarios where there are a large number of mixed discrete and continuous auxiliary variables. There exist methodologies developed in the missing data and causal inference literature that could be adapted for non-probability samples. Section 5 of Wu (2022) contains a brief discussion on rank-based methods. The rank-based method described in Beaumont et al. (2024) has similar spirits. This is an important topic that requires further research.

Variable selection using AIC or other similar criteria requires a true likelihood function. Beaumont et al. (2024) demonstrated the usefulness of the pseudo likelihood function of Chen et al. (2020) in the context of variable selection. I am excited to see such development and look forward to seeing more research effort in that direction.

Acknowledgements

This research was supported by grants from the Natural Sciences and Engineering Research Council of Canada and the Canadian Statistical Sciences Institute.

References

Beaumont, J.-F., Bosa, K., Brennan, A., Charlebois, J. and Chu, K. (2024). Handling non-probability samples through inverse probability weighting with an application to Statistics Canada's crowdsourcing data. *Survey Methodology*, 50, 1, 77-106. Paper available at http://www.statcan.gc.ca/pub/12-001-x/2024001/article/00004-eng.pdf.

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

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

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.

Wu, C. (2022). Statistical inference with non-probability survey samples (with Discussion). *Survey Methodology*, 48, 2, 283-311. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2022002/article/00002-eng.pdf.

Survey Methodology, June 2024
Vol. 50, No. 1, pp. 123-141
Statistics Canada, Catalogue No. 12-001-X

Authors’ response to comments on “Handling non-probability samples through inverse probability weighting with an application to Statistics Canada’s crowdsourcing data”:

Some new developments on likelihood approaches to estimation of participation probabilities for non-probability samples

Jean-François Beaumont, Keven Bosa, Andrew Brennan, Joanne Charlebois and Kenneth Chu[1]

Abstract

Inspired by the two excellent discussions of our paper, we offer some new insights and developments into the problem of estimating participation probabilities for non-probability samples. First, we propose an improvement of the method of Chen, Li and Wu (2020), based on best linear unbiased estimation theory, that more efficiently leverages the available probability and non-probability sample data. We also develop a sample likelihood approach, similar in spirit to the method of Elliott (2009), that properly accounts for the overlap between both samples when it can be identified in at least one of the samples. We use best linear unbiased prediction theory to handle the scenario where the overlap is unknown. Interestingly, our two proposed approaches coincide in the case of unknown overlap. Then, we show that many existing methods can be obtained as a special case of a general unbiased estimating function. Finally, we conclude with some comments on nonparametric estimation of participation probabilities.

Key Words: Best linear unbiased estimation; Best linear unbiased prediction; Estimating equation; Population likelihood; Pseudo likelihood; Sample likelihood.

1. General remarks

We would first like to thank the discussants for taking the time to read our paper and share their thoughtful and insightful observations on inverse probability weighting for non-probability samples. Dr. Wu shed light into three common weighting methods for non-probability samples whereas Dr. Gershunskaya and Dr. Beresovsky introduced two new methods to us: Implicit Logistic Regression (ILR), see also Beresovsky (2019), Savitsky, Williams, Gershunskaya, Beresovsky and Johnson (2022) and Gershunskaya and Lahiri (2023), and Pseudo-ILR. We have greatly enjoyed reading both discussions, which have helped us improve our knowledge on the field and stimulated us to have further thoughts on the topic. In what follows, we will share these thoughts along with some new developments.

Sections 2, 3 and 4 are devoted to the methods of Chen, Li and Wu (2020), Wang, Valliant and Li (2021) and Elliott (2009), see also Elliott and Valliant (2017), respectively. We provide additional observations on

[1] Jean-François Beaumont, Keven Bosa, Andrew Brennan, Joanne Charlebois and Kenneth Chu, Statistics Canada, 150 Tunney’s Pasture Driveway, Ottawa, Ontario, K1A 0T6. E-mail: jean-francois.beaumont@statcan.gc.ca, keven.bosa@statcan.gc.ca, andrew.brennan@statcan.gc.ca, joanne.charlebois@statcan.gc.ca and kenneth.chu@statcan.gc.ca.

these methods and make connections with the ILR and Pseudo-ILR methods. We show that all three methods are valid in the sense that they lead to unbiased estimating functions for the parameters of the participation model, regardless of the size of the probability and non-probability samples as well as the size of the overlap between both samples. However, only the Chen-Li-Wu (CLW) method has the property of reducing to the maximum likelihood method when the probability sample is a census, which we refer to as the Census Likelihood (CL) property. In Section 2, we also show that the CLW method does not fully leverage the available auxiliary information, which may result in an inefficient estimating function, particularly when the non-probability sample is larger than the probability sample. Using Best Linear Unbiased (BLU) estimation theory, we propose an improvement of the CLW method that addresses this issue and still satisfies the CL property. In Section 5, we propose a sample likelihood approach, similar in spirit to the Elliott/ILR method, that properly accounts for the overlap between both samples provided that it can be identified in one of the samples. Our sample likelihood approach satisfies the CL property. Using BLU prediction theory, we obtain an "optimal" estimating function applicable when the overlap cannot be identified in any of the samples. Interestingly, it is identical to the estimating function underlying our improved CLW method. In Section 6, we unify existing methods that do not require identifying the overlap and show their equivalence for the homogeneous group model. A brief summary is given in Section 7 along with a few comments on nonparametric estimation of participation probabilities.

2. A population likelihood approach and the method of Chen, Li and Wu (2020)

We use the notation in our main paper: the vector of auxiliary variables for unit k of the finite population U is denoted by $\mathbf{x}_k$, and the participation indicator (indicator of participation in the non-probability sample $s_{\text{NP}} \subset U$) for population unit $k \in U$ is denoted by δ_k. The participation probability $p_k = \Pr\left(\delta_k = 1 \,|\, \mathbf{x}_k\right) > 0$ is modelled using a parametric model, such as the logistic model $p_k(\boldsymbol{\alpha}) = \left[1 + \exp\left(-\mathbf{x}_k' \boldsymbol{\alpha}\right)\right]^{-1}$, where $\boldsymbol{\alpha}$ is a vector of unknown model parameters. We make the following standard independence assumption:

A1) $\delta_k,\ k \in U,$ are mutually independent given $\mathbf{x}_k,\ k \in U.$

Ideally, we would have access to both $\left\{\mathbf{x}_k ; k \in s_{\text{NP}}\right\}$ and $\left\{\mathbf{x}_k ; k \in U\right\}$. These two data sets would not need to be linked. Under that ideal scenario and assumption (A1), the population log likelihood function is

$$l(\boldsymbol{\alpha}) = \log\left\{\prod_{k \in U}\left[p_k(\boldsymbol{\alpha})\right]^{\delta_k}\left[1 - p_k(\boldsymbol{\alpha})\right]^{(1-\delta_k)}\right\} = \sum\nolimits_{k \in s_{\text{NP}}} \log\left[\frac{p_k(\boldsymbol{\alpha})}{1 - p_k(\boldsymbol{\alpha})}\right] + \sum\nolimits_{k \in U} \log\left[1 - p_k(\boldsymbol{\alpha})\right]$$

and the population likelihood estimating function (or score function) is

$$\mathbf{U}(\boldsymbol{\alpha}) = \sum\nolimits_{k \in s_{\text{NP}}} \frac{1}{p_k(\boldsymbol{\alpha})} \frac{\mathbf{g}_k(\boldsymbol{\alpha})}{\left[1 - p_k(\boldsymbol{\alpha})\right]} - \sum\nolimits_{k \in U} \frac{\mathbf{g}_k(\boldsymbol{\alpha})}{\left[1 - p_k(\boldsymbol{\alpha})\right]}, \tag{2.1}$$

where $\mathbf{g}_k(\boldsymbol{\alpha}) = \partial p_k(\boldsymbol{\alpha}) / \partial \boldsymbol{\alpha}$. In particular, $\mathbf{g}_k(\boldsymbol{\alpha}) = p_k(\boldsymbol{\alpha})\left[1 - p_k(\boldsymbol{\alpha})\right]\mathbf{x}_k$ for the logistic model.

In many real cases, the vector $\mathbf{x}_k$ is not known for the entire population, but is at least available in a probability sample s_P in addition to the non-probability sample s_{NP}. As a result, the term $\sum_{k \in U} \log\left[1 - p_k(\boldsymbol{\alpha})\right]$ in the population log likelihood function $l(\boldsymbol{\alpha})$ is not computable as it depends on unknown values of $\mathbf{x}_k$. Chen, Li and Wu (2020) proposed to address this issue by estimating this term using the probability sample. This leads to the pseudo log likelihood function:

$$\hat{l}(\boldsymbol{\alpha}) = \sum_{k \in s_{\mathrm{NP}}} \log\left[\frac{p_k(\boldsymbol{\alpha})}{1 - p_k(\boldsymbol{\alpha})}\right] + \sum_{k \in s_P} w_k \log\left[1 - p_k(\boldsymbol{\alpha})\right], \tag{2.2}$$

where $w_k = 1/\pi_k$ is a probability survey weight for unit $k \in s_P$ and π_k is its selection probability. We focus on this basic weight for simplicity although more complex weighting methods, involving nonresponse and calibration adjustments, are often used in real surveys. Taking the derivative of (2.2) with respect to $\boldsymbol{\alpha}$, we obtain the pseudo likelihood estimating function

$$\hat{\mathbf{U}}_\pi(\boldsymbol{\alpha}) = \sum_{k \in s_{\mathrm{NP}}} \frac{1}{p_k(\boldsymbol{\alpha})} \frac{\mathbf{g}_k(\boldsymbol{\alpha})}{\left[1 - p_k(\boldsymbol{\alpha})\right]} - \sum_{k \in s_P} w_k \frac{\mathbf{g}_k(\boldsymbol{\alpha})}{\left[1 - p_k(\boldsymbol{\alpha})\right]}. \tag{2.3}$$

It is easy to see that the estimating function (2.3) is md unbiased, conditional on π_k and $\mathbf{x}_k$, $k \in U$, i.e., $E_{md}\left[\hat{\mathbf{U}}_\pi(\boldsymbol{\alpha})\right] = \mathbf{0}$, provided that the following assumption holds:

A2) $E_m\left(\delta_k \middle| \pi_k, \mathbf{x}_k\right) = E_m\left(\delta_k \middle| \mathbf{x}_k\right) = p_k(\boldsymbol{\alpha}),\ k \in U.$

The subscript m refers to the participation model and the subscript d refers to the probability sampling design. Conditioning on π_k, $k \in U$, makes sense when π_k is available in both samples since it can be treated as a potential auxiliary variable. Indeed, assumption (A2) is automatically satisfied if π_k is included in the vector $\mathbf{x}_k$. In Section 2.1, we will condition on π_k, $k \in U$. Then, in Section 2.2, we will consider the case where π_k is treated as random and inferences are conditional only on $\mathbf{x}_k$, $k \in U$.

2.1 Improvement of the CLW estimating function using BLU estimation theory

The second term on the right-hand side of (2.3) is an estimator of the corresponding term in (2.1), $\boldsymbol{\Phi}(\boldsymbol{\alpha}) = \sum_{k \in U} \mathbf{g}_k(\boldsymbol{\alpha}) / \left[1 - p_k(\boldsymbol{\alpha})\right]$. It is an inefficient estimator of $\boldsymbol{\Phi}(\boldsymbol{\alpha})$ because it uses only probability sample data and ignores relevant non-probability sample auxiliary data. A more efficient estimator would thus use auxiliary data from both samples. Such an estimator could be obtained by applying the Missing Information Principle (MIP). The MIP was introduced by Orchard and Woodbury (1972), see also Chambers (2023) for a recent reference on applications of the MIP with survey data. The MIP consists of replacing the population likelihood estimating function (2.1) with its expectation conditional on observed data, or equivalently replacing $\boldsymbol{\Phi}(\boldsymbol{\alpha})$ with its best predictor. However, this would involve modelling the vector of auxiliary variables, and the MIP solution would generally not be easy to implement.

As an alternative to applying the MIP, we propose to estimate $\boldsymbol{\Phi}(\boldsymbol{\alpha})$ using BLU estimation theory. We consider the following linear unbiased estimator that uses auxiliary data from both samples:

$$\hat{\boldsymbol{\Phi}}(\boldsymbol{\alpha}) = \sum\nolimits_{k \in s_{\mathrm{NP}}} \frac{\gamma_k}{p_k(\boldsymbol{\alpha})} \frac{\mathbf{g}_k(\boldsymbol{\alpha})}{\left[1 - p_k(\boldsymbol{\alpha})\right]} + \sum\nolimits_{k \in s_P} w_k (1 - \gamma_k) \frac{\mathbf{g}_k(\boldsymbol{\alpha})}{\left[1 - p_k(\boldsymbol{\alpha})\right]}, \tag{2.4}$$

where $\gamma_k,\, k \in U,$ are constants. It is easy to show that $\hat{\boldsymbol{\Phi}}(\boldsymbol{\alpha})$ is md unbiased for $\boldsymbol{\Phi}(\boldsymbol{\alpha}),$ i.e., $E_{md}\left[\hat{\boldsymbol{\Phi}}(\boldsymbol{\alpha})\right] = \boldsymbol{\Phi}(\boldsymbol{\alpha}),$ provided that assumption (A2) holds. Replacing the second term on the right-hand side of (2.1) with the right-hand side of (2.4), we obtain the md unbiased estimating function

$$\hat{\mathbf{U}}_{\pi}^{\gamma}(\boldsymbol{\alpha}) = \sum\nolimits_{k \in s_{\mathrm{NP}}} \frac{1}{p_k(\boldsymbol{\alpha})} \frac{(1 - \gamma_k)}{\left[1 - p_k(\boldsymbol{\alpha})\right]} \mathbf{g}_k(\boldsymbol{\alpha}) - \sum\nolimits_{k \in s_P} w_k \frac{(1 - \gamma_k)}{\left[1 - p_k(\boldsymbol{\alpha})\right]} \mathbf{g}_k(\boldsymbol{\alpha}). \tag{2.5}$$

It is easy to see that the CLW estimating function (2.3) is the special case of (2.5) obtained by specifying $\gamma_k = 0$ for all $k \in U.$

The BLU estimator of $\boldsymbol{\Phi}(\boldsymbol{\alpha})$ is obtained by finding $\gamma_k,\, k \in U,$ that minimize $\operatorname{var}_{md}\left[\mathbf{c}'\hat{\boldsymbol{\Phi}}(\boldsymbol{\alpha})\right]$ for any fixed vector $\mathbf{c} \neq \mathbf{0}.$ We make the following assumptions:

A3) $I_k,\, k \in U,$ are mutually independent given π_k and $\mathbf{x}_k,\, k \in U,$ where I_k is the indicator of inclusion in the probability sample $s_P.$

A4) $E_d\left(I_k \left| \delta_k, \pi_k, \mathbf{x}_k \right.\right) = E_d\left(I_k \left| \pi_k, \mathbf{x}_k \right.\right) = \pi_k,\, k \in U.$

Assumption (A3) implies that the probability sample is selected using Poisson sampling. It is used to simplify the derivations of $\operatorname{var}_{md}\left[\mathbf{c}'\hat{\boldsymbol{\Phi}}(\boldsymbol{\alpha})\right]$ even though we recognize that other sampling designs may be used in practice. Note that neither assumption (A3) nor assumptions (A1) and (A4) are needed to prove that the estimating function (2.5) is md unbiased. Under (A1)-(A4), it is straightforward to show that

$$\operatorname{var}_{md}\left[\mathbf{c}'\hat{\boldsymbol{\Phi}}(\boldsymbol{\alpha})\right] = \sum\nolimits_{k \in U} \frac{\left[1 - p_k(\boldsymbol{\alpha})\right]}{p_k(\boldsymbol{\alpha})} \gamma_k^2 \left(\frac{\mathbf{c}'\mathbf{g}_k(\boldsymbol{\alpha})}{1 - p_k(\boldsymbol{\alpha})}\right)^2 + \sum\nolimits_{k \in U} \frac{(1 - \pi_k)}{\pi_k} (1 - \gamma_k)^2 \left(\frac{\mathbf{c}'\mathbf{g}_k(\boldsymbol{\alpha})}{1 - p_k(\boldsymbol{\alpha})}\right)^2. \tag{2.6}$$

The variance (2.6) is minimized when

$$\gamma_k = \gamma_{\pi,k}^{\mathrm{opt}}(\boldsymbol{\alpha}) = \frac{(1 - \pi_k) p_k(\boldsymbol{\alpha})}{(1 - \pi_k) p_k(\boldsymbol{\alpha}) + \left[1 - p_k(\boldsymbol{\alpha})\right] \pi_k} = \frac{w_k - 1}{(w_k - 1) + (p_k^{-1}(\boldsymbol{\alpha}) - 1)},\ k \in U. \tag{2.7}$$

Plugging $\gamma_{\pi,k}^{\mathrm{opt}}(\boldsymbol{\alpha})$ in (2.4) leads to the BLU estimator of $\boldsymbol{\Phi}(\boldsymbol{\alpha}).$

We have the following properties associated with $\gamma_{\pi,k}^{\mathrm{opt}}(\boldsymbol{\alpha})$:

i) If $\pi_k = p_k(\boldsymbol{\alpha})$ then $\gamma_{\pi,k}^{\mathrm{opt}}(\boldsymbol{\alpha}) = 1/2;$

ii) If $\pi_k > p_k(\boldsymbol{\alpha})$ then $0 \leq \gamma_{\pi,k}^{\mathrm{opt}}(\boldsymbol{\alpha}) < 1/2;$

iii) If $\pi_k < p_k(\boldsymbol{\alpha})$ then $1/2 < \gamma_{\pi,k}^{\mathrm{opt}}(\boldsymbol{\alpha}) \leq 1;$

iv) If $\pi_k \to 1$ or $p_k(\boldsymbol{\alpha}) \to 0$ then $\gamma_{\pi,k}^{\mathrm{opt}}(\boldsymbol{\alpha}) \to 0;$ and

v) If $\pi_k \to 0$ or $p_k(\boldsymbol{\alpha}) \to 1$ then $\gamma_{\pi,k}^{\mathrm{opt}}(\boldsymbol{\alpha}) \to 1.$

As a result of properties (iii) and (v), when the probability sample is small compared with the non-probability sample, $\gamma_{\pi,k}^{\text{opt}}(\boldsymbol{\alpha})$ is expected to be large for many population units, and the CLW method ($\gamma_k = 0$) may become inefficient relative to the optimal solution (2.7). The inefficiency of the CLW method in that scenario was shown in the empirical study of Savitsky et al. (2022). The explanation for this inefficiency is that the CLW method ignores the large non-probability sample for the estimation of the population total $\boldsymbol{\Phi}(\boldsymbol{\alpha})$. From properties (ii) and (iv), the CLW method should perform better in the reverse scenario where the probability sample is much larger than the non-probability sample as it possesses the CL property, i.e., the estimating function (2.3) reduces to the population likelihood estimating function (2.1) when the probability sample is a census. This scenario is not unrealistic in practice. For example, the Canadian Long-Form Census, randomly administered to 25% of the Canadian population, could be an effective probability sample for the estimation of participation probabilities for a smaller non-probability sample.

If we plug (2.7) in the estimating function (2.5), we obtain the "optimal" estimating function:

$$\begin{aligned}\hat{\mathbf{U}}_{\pi}^{\text{opt}}(\boldsymbol{\alpha}) &= \sum\nolimits_{k \in s_{\text{NP}}} \frac{1}{p_k(\boldsymbol{\alpha})} \frac{\pi_k}{\left[\pi_k + p_k(\boldsymbol{\alpha}) - 2\pi_k p_k(\boldsymbol{\alpha})\right]} \mathbf{g}_k(\boldsymbol{\alpha}) \\ &\quad - \sum\nolimits_{k \in s_P} \frac{1}{\pi_k} \frac{\pi_k}{\left[\pi_k + p_k(\boldsymbol{\alpha}) - 2\pi_k p_k(\boldsymbol{\alpha})\right]} \mathbf{g}_k(\boldsymbol{\alpha}).\end{aligned} \tag{2.8}$$

It is straightforward to show that $\hat{\mathbf{U}}_{\pi}^{\text{opt}}(\boldsymbol{\alpha})$ is the BLU predictor of $\mathbf{U}(\boldsymbol{\alpha})$ given in (2.1). Like the CLW estimating function (2.3), it possesses the CL property.

2.2 Weight smoothing

As discussed above, the possible inefficiency of (2.3) can be mostly explained by the omission of relevant non-probability sample auxiliary data for the estimation of $\boldsymbol{\Phi}(\boldsymbol{\alpha})$. Another possible source of inefficiency may be attributable to the variability of the survey weights w_k, $k \in s_P$. Indeed, it is well known that pseudo likelihood estimation can be inefficient for the estimation of model parameters from probability survey data (e.g., see Chambers, 2023, for a recent reference). Weight smoothing (Beaumont, 2008) can be used to address this issue. In this context, it consists of replacing the survey weight w_k in (2.3) with the smoothed weight $\tilde{w}_k = E_{\xi}\left(w_k \mid k \in s_P, \mathbf{x}_k\right)$, where the subscript ξ indicates that the expectation is taken with respect to a model for π_k (or w_k). The smoothed weight $\tilde{w}_k$ is often unknown but can be estimated using the probability sample along with parametric or nonparametric models. If π_k is available in the non-probability sample and included in the vector $\mathbf{x}_k$, $\tilde{w}_k = w_k$ and weight smoothing does not bring any efficiency gain.

Using a relationship in Pfeffermann and Sverchkov (1999), the smoothed weight can be expressed as

$$\tilde{w}_k = E_{\xi}\left(w_k \mid k \in s_P, \mathbf{x}_k\right) = \frac{1}{E_{\xi}\left(\pi_k \mid \mathbf{x}_k\right)} = \frac{1}{\tilde{\pi}_k}, \tag{2.9}$$

where $\tilde{\pi}_k = E_\xi\left(\pi_k \,|\, \mathbf{x}_k\right) = \Pr\left(k \in s_P \,|\, \mathbf{x}_k\right)$. Using this relationship, it is straightforward to show that the estimating function (2.3) is ξmd unbiased, regardless of the validity of assumption (A2) and whether w_k or $\tilde{w}_k$ is used in (2.3), i.e., $E_{\xi md}\left[\hat{\mathbf{U}}_\pi(\boldsymbol{\alpha})\right] = \mathbf{0}$ and $E_{\xi md}\left[\hat{\mathbf{U}}_{\tilde{\pi}}(\boldsymbol{\alpha})\right] = \mathbf{0}$, where $\hat{\mathbf{U}}_{\tilde{\pi}}(\boldsymbol{\alpha})$ is the estimating function (2.3) with w_k replaced by $\tilde{w}_k$. Note that ξmd expectations are conditional only on $\mathbf{x}_k$, $k \in U$, so that π_k is treated as random. Relationship (2.9) can also be used to obtain an estimator of $\tilde{\pi}_k$, i.e., a consistent estimator of $\tilde{\pi}_k = E_\xi\left(\pi_k \,|\, \mathbf{x}_k\right)$ is $\hat{\tilde{\pi}}_k = 1/\hat{\tilde{w}}_k$, where $\hat{\tilde{w}}_k$ is a consistent estimator of $\tilde{w}_k = E_\xi\left(w_k \,|\, k \in s_P, \mathbf{x}_k\right)$.

Using $\tilde{w}_k$ instead of w_k in the estimating function (2.3) increases its efficiency at the expense of requiring the validity of a model for w_k and estimation of $\tilde{w}_k$. A similar argument can be made to improve the efficiency of $\hat{\boldsymbol{\Phi}}(\boldsymbol{\alpha})$ by replacing w_k by $\tilde{w}_k$ in (2.4). The resulting estimator is ξmd unbiased, i.e., $E_{\xi md}\left[\hat{\boldsymbol{\Phi}}(\boldsymbol{\alpha})\right] = \boldsymbol{\Phi}(\boldsymbol{\alpha})$, and its variance $\operatorname{var}_{\xi md}\left[\mathbf{c}'\hat{\boldsymbol{\Phi}}(\boldsymbol{\alpha})\right]$ takes the same form as (2.6), with π_k replaced by $\tilde{\pi}_k$, provided that assumptions (A1)-(A4) hold as well as the following assumption:

A5) π_k, $k \in U$, are mutually independent given $\mathbf{x}_k$, $k \in U$.

As a result, the optimal value of γ_k, denoted by $\gamma_{\tilde{\pi},k}^{\text{opt}}(\boldsymbol{\alpha})$, and the optimal estimating function, denoted by $\hat{\mathbf{U}}_{\tilde{\pi}}^{\text{opt}}(\boldsymbol{\alpha})$, are again given by expressions (2.7) and (2.8), respectively, with π_k replaced by $\tilde{\pi}_k$.

Using π_k in (2.8) is not possible if it is not observed in the non-probability sample, a likely scenario in practice. In that case, an estimate of $\tilde{\pi}_k$ can be used in (2.8) to replace π_k. If π_k is observed in the non-probability sample but not included in $\mathbf{x}_k$, it may still be desirable to replace π_k with an estimate of $\tilde{\pi}_k$ to improve the efficiency of the optimal estimating function (2.8).

2.3 Variable selection

The estimating function (2.8) is not the derivative of a pseudo log likelihood function. Therefore, the methodology that we used in our main paper to derive an Akaike Information Criterion (AIC), based on Lumley and Scott (2015), is not directly applicable. For variable selection, one option is to use our proposed AIC along with the CLW method. Once auxiliary variables are selected, the estimating function (2.8), $\hat{\mathbf{U}}_\pi^{\text{opt}}(\boldsymbol{\alpha})$, or its smoothed version $\hat{\mathbf{U}}_{\tilde{\pi}}^{\text{opt}}(\boldsymbol{\alpha})$, could be used to estimate $\boldsymbol{\alpha}$ instead of the CLW estimating function. Otherwise, variable selection methods that are not likelihood-based could be envisioned.

As a side remark, the methodology developed in this section to obtain an optimal estimator (or BLU estimator) of $\boldsymbol{\Phi}(\boldsymbol{\alpha})$ could also be used to combine two independent probability samples from the same population. This is left for future research.

3. The method of Wang, Valliant and Li (2021) and Pseudo-ILR

The method of Wang-Valliant-Li (WVL) consists of creating an artificial population U_A by stacking the non-probability sample s_{NP} on top of the population U. Each element of U_A is considered distinct even though non-probability sample units are present twice in U_A. An indicator R_i is defined for each element

$i \in U_A$; $R_i = 1$, if $i \in s_{\text{NP}} \cap U_A$, and $R_i = 0$, if $i \in U \cap U_A$. We use the subscript i to refer to elements of the artificial population U_A so as to distinguish them from the units of the population U. For any given unit $k \in s_{\text{NP}}$, there are two distinct elements of U_A that are labelled differently from that unit k; $R = 0$ for one element and $R = 1$ for the other. The probability sample is assumed to be selected from the elements in $U \cap U_A$ for which $R_i = 0$. The authors also assumed that the indicators R_i, $i \in U_A$, are mutually independent given $\mathbf{x}_i$, $i \in U_A$, and obtained a pseudo log likelihood function similar to CLW by modelling $\Pr\left(R_i = 1 \middle| i \in U_A, \mathbf{x}_i\right)$ using a logistic model. Then, they established the relationship $\Pr\left(R_i = 1 \middle| i \in U_A, \mathbf{x}_i\right) = p_i/(1+p_i)$, which allowed them to estimate the participation probability p_i. Because they used a logistic model for $\Pr\left(R_i = 1 \middle| i \in U_A, \mathbf{x}_i\right)$, they implicitly modelled p_i using the exponential model $p_i(\boldsymbol{\alpha}) = \exp\left(\mathbf{x}_i'\boldsymbol{\alpha}\right)$, which has the undesirable feature of admitting estimates greater than 1. However, nothing in their theory would have prevented them from using another model for p_i, such as the logistic model, and thereby implicitly obtaining a model for $\Pr\left(R_i = 1 \middle| i \in U_A, \mathbf{x}_i\right)$. This is exactly what Dr. Gershunskaya and Dr. Beresovsky proposed in their discussion. They call their method Pseudo-ILR, which is simply the WVL method with a logistic model for p_i.

For an arbitrary parametric model for p_i, the WVL or Pseudo-ILR estimating function can be expressed as

$$\hat{\mathbf{U}}_{\pi}^{\text{WVL-PILR}}(\boldsymbol{\alpha}) = \sum\nolimits_{k \in s_{\text{NP}}} \frac{1}{p_k(\boldsymbol{\alpha})} \frac{\mathbf{g}_k(\boldsymbol{\alpha})}{\left[1 + p_k(\boldsymbol{\alpha})\right]} - \sum\nolimits_{k \in s_P} w_k \frac{\mathbf{g}_k(\boldsymbol{\alpha})}{\left[1 + p_k(\boldsymbol{\alpha})\right]}. \tag{3.1}$$

Similar to the CLW method, the estimating function (3.1) can possibly be improved by replacing the survey weight w_k with the smoothed weight $\tilde{w}_k$. The resulting estimating function is denoted by $\hat{\mathbf{U}}_{\tilde{\pi}}^{\text{WVL-PILR}}(\boldsymbol{\alpha})$.

Dr. Gershunskaya and Dr. Beresovsky pointed out that, unlike δ_k, the indicator R_i is fully observed once the probability and non-probability samples are observed. However, this characteristic of R_i is deceiving. The indicators R_i for elements in the probability and non-probability samples do not bring any new information about the participation mechanism than what is observable about δ_k and used in the CLW method. In other words, both methods use the same observed information: $\left\{w_k, \mathbf{x}_k; k \in s_P\right\}$ and $\left\{\delta_k, \mathbf{x}_k; k \in s_{\text{NP}}\right\}$. In addition, we see two main issues with the WVL method, which are described below.

Issue 1: The assumption that R_i, $i \in U_A$, are mutually independent given $\mathbf{x}_i$, $i \in U_A$, is not valid as each non-probability sample unit is present twice in U_A: $R_i = 0$ for one element and $R_i = 1$ for the other (see greater detail below).

Issue 2: The estimating function (3.1) does not have the CL property as it does not reduce to the population likelihood estimating function (2.1) when the probability sample is a census. When $s_P = U$, the CLW method uses the same amount of information as if $\left\{\delta_k, \mathbf{x}_k; k \in U\right\}$ were known, but the WVL log likelihood fails to recognize this information.

Despite the above two issues, it is easy to show that the estimating function (3.1) is md unbiased provided that assumption (A2) holds. Both $\hat{\mathbf{U}}_{\pi}^{\text{WVL-PILR}}(\boldsymbol{\alpha})$ and $\hat{\mathbf{U}}_{\tilde{\pi}}^{\text{WVL-PILR}}(\boldsymbol{\alpha})$ are also ξmd unbiased, regardless of the validity of assumption (A2); the estimating function (3.1) is thus valid. It can be written in the form (2.5) with $\gamma_k = \gamma_k^{\text{WVL-PILR}}(\boldsymbol{\alpha}) = 2\, p_k(\boldsymbol{\alpha}) \big/ \left[1 + p_k(\boldsymbol{\alpha})\right].$ It is easy to see that $0 < \gamma_k^{\text{WVL-PILR}}(\boldsymbol{\alpha}) \leq 1;$ it thus uses non-probability sample auxiliary data to some extent for the estimation of $\boldsymbol{\Phi}(\boldsymbol{\alpha})$.

The variance (2.6) is a quadratic function of γ_k that is minimized when $\gamma_k = \gamma_{\pi,k}^{\text{opt}}(\boldsymbol{\alpha}),\ k \in U.$ Therefore, if $\gamma_{\pi,k}^{\text{opt}}(\boldsymbol{\alpha})$ is closer to $\gamma_k^{\text{WVL-PILR}}(\boldsymbol{\alpha})$ than to 0 for most population units, i.e., $\gamma_{\pi,k}^{\text{opt}}(\boldsymbol{\alpha}) > 0.5\gamma_k^{\text{WVL-PILR}}(\boldsymbol{\alpha}),$ the WVL estimating function (3.1) is expected to be more efficient than the CLW estimating function ($\gamma_k = 0$), but still less efficient than the optimal estimating function (2.8). It is easy to show that $\gamma_{\pi,k}^{\text{opt}}(\boldsymbol{\alpha}) > 0.5\gamma_k^{\text{WVL-PILR}}(\boldsymbol{\alpha})$ is satisfied when $\pi_k < \left[2 - p_k(\boldsymbol{\alpha})\right]^{-1}.$ However, if $\pi_k > \left[2 - p_k(\boldsymbol{\alpha})\right]^{-1}$ for most population units, the CLW estimating function (2.3) is expected to become more efficient than (3.1), in particular when the probability sample is a census ($\pi_k = 1, k \in U$). Since $\left[2 - p_k(\boldsymbol{\alpha})\right]^{-1} > 0.5,$ the condition $\pi_k < \left[2 - p_k(\boldsymbol{\alpha})\right]^{-1}$ is typically more common in social surveys than $\pi_k > \left[2 - p_k(\boldsymbol{\alpha})\right]^{-1}$, at least for most population units. It is also straightforward to show that $\gamma_{\pi,k}^{\text{opt}}(\boldsymbol{\alpha}) = \gamma_k^{\text{WVL-PILR}}(\boldsymbol{\alpha})$ when $\pi_k = 1/3$ whereas $\gamma_{\pi,k}^{\text{opt}}(\boldsymbol{\alpha}) = 0$ when $\pi_k = 1.$ The WVL estimating function should thus be close to the optimal estimating function when the probability sample size is around one third of the population size and the selection probabilities π_k are not too variable.

Like Prof. Wu in his discussion, we had trouble understanding the probabilistic framework underlying the relationship $\Pr\left(R_i = 1 \middle| i \in U_A, \mathbf{x}_i\right) = p_i / (1 + p_i).$ However, it appears that the clever setup proposed by Savitsky et al. (2022) provides a correct justification of this relationship. These authors imagined a fixed augmented population U^* obtained by stacking population U_1 on top of population U_2, where U_1 and U_2 are two populations identical to U of size N, but uniquely labelled so that each of the $2N$ elements of U^* is viewed as distinct. First, one of the two populations is chosen at random with probability ½. We denote that randomly selected population by U_{NP}, which is either U_1 or U_2. The other population is denoted by U_P. The non-probability sample s_{NP} is observed from U_{NP}, and the probability sample s_P is randomly selected from U_P. Using this setup, it is easy to show that

$$\Pr\left(R_i = 1 \middle| i \in s_{\text{NP}} \cup U_P, \mathbf{x}_i\right) = \frac{\Pr\left(i \in s_{\text{NP}} \middle| \mathbf{x}_i\right)}{\Pr\left(i \in s_{\text{NP}} \cup U_P \middle| \mathbf{x}_i\right)} = \frac{1/2\, p_i}{1/2 p_i + 1/2} = \frac{p_i}{(1 + p_i)}.$$

Note that the random splitting of U^* into U_{NP} and U_P is not explicitly stated in Savitsky et al. (2022) but is necessary to obtain the above equation.

This setup seems to solve the problem but the two issues noted above remain. In particular, it is easy to show that the independence assumption is not satisfied since, for any pair of elements $i \in U_1$ and $j \in U_2$,

$$\Pr\left(R_i = 1, R_j = 1 \middle| i, j \in s_{\text{NP}} \cup U_P, \mathbf{x}_i, \mathbf{x}_j\right) = 0 \neq \frac{p_i}{(1 + p_i)} \frac{p_j}{(1 + p_j)}.$$

For two different elements i and j in the same population, either U_1 or U_2, we have

$$\Pr\left(R_i = 1, R_j = 1 \middle| i, j \in s_{\text{NP}} \cup U_P, \mathbf{x}_i, \mathbf{x}_j\right) = \frac{p_i p_j}{1 + p_i p_j} \neq \frac{p_i}{(1 + p_i)} \frac{p_j}{(1 + p_j)}.$$

provided (A1) holds for elements $i \in U_{\mathrm{NP}}$. Therefore, the independence assumption is reasonable only when all (or at least many of) the participation probabilities are small, and thereby the overlap is a small portion of the probability sample. In this situation, the WVL and CWL estimating functions should be roughly equivalent. If most of the participation probabilities are large, the pseudo log likelihood function proposed by WVL is based on an incorrect independence assumption. In principle, an AIC based on an incorrect (pseudo) log likelihood function is not valid. How small should the participation probabilities be to make this independence assumption reasonable? The simulation study of Dr. Gershunskaya and Dr. Beresovsky is a first step in that direction, but further studies are needed. Note that the CLW pseudo log likelihood function is valid regardless of the magnitude of the participation probabilities as long as assumption (A1) holds.

4. The method of Elliott (2009) and ILR

In the method of Elliott (2009), see also Elliott and Valliant (2017), a combined sample s^* is obtained by stacking the non-probability sample s_{NP} on top of the probability sample s_P while ignoring the possible (unknown) overlap. A population unit $k \in U$ that is selected in s_P and observed in s_{NP} is thus present twice in s^*. Elliott (2009) implicitly assumed that the overlap between both samples is negligible. Similar to Wang, Valliant and Li (2021), an indicator z_i, $i \in s^*$, is created such that $z_i = 1$, if $i \in s_{\mathrm{NP}} \cap s^*$, and $z_i = 0$, if $i \in s_P \cap s^*$. Elliott (2009) proposed to model $\rho_i = \Pr\left(z_i = 1 \middle| i \in s^*, \mathbf{x}_i\right)$ using a logistic model and, assuming the sampling fractions are small (Elliott and Valliant, 2017), established the relationship $p_i = K\tilde{\pi}_i \, \rho_i / (1 - \rho_i)$ used to estimate p_i, where K is an unknown constant of proportionality. This implies that $\rho_i = p_i / (K\tilde{\pi}_i + p_i)$. In practice, $\tilde{\pi}_i$ is typically unknown but can be estimated, as discussed in Section 2.

In this section and the next one, we condition on $\mathbf{x}_i$ and treat π_i as random. The theory remains valid if we condition on both $\mathbf{x}_i$ and π_i provided that $\tilde{\pi}_i$ is replaced with π_i in the developments below and assumption (A2) holds. Conditioning on π_i makes sense only if it is observed in both samples, so that it can be treated as a potential auxiliary variable and included in $\mathbf{x}_i$. If π_i is included in $\mathbf{x}_i$, $\tilde{\pi}_i = \pi_i$ and assumption (A2) is satisfied. For complex probability surveys, it is unlikely that π_i would be observed in the non-probability sample. In that case, it must be treated as random.

Using the setup introduced by Savitsky et al. (2022), also described in Section 3, it is easy to show that

$$\rho_i = \Pr\left(z_i = 1 \middle| i \in s^*, \mathbf{x}_i\right) = \frac{\Pr\left(i \in s_{\mathrm{NP}} \middle| \mathbf{x}_i\right)}{\Pr\left(i \in s^* \middle| \mathbf{x}_i\right)} = \frac{p_i}{(\tilde{\pi}_i + p_i)}.$$

Note that the relationship does not require a constant of proportionality $(K = 1)$ and is valid regardless of the size of the sampling fractions. When the logistic model $\rho_i(\boldsymbol{\alpha}) = \left[1 + \exp\left(-\mathbf{x}_i'\boldsymbol{\alpha}\right)\right]^{-1}$ is used, it is easy to see that the resulting implicit model for p_i is $p_i(\boldsymbol{\alpha}) = \tilde{\pi}_i \exp\left(\mathbf{x}_i'\boldsymbol{\alpha}\right)$, which admits estimates greater than 1. Other models for p_i can be considered, such as the logistic model. Beresovsky (2019), see also Gershunskaya and Lahiri (2023), called this method ILR, which is essentially the Elliott method with a logistic model for p_i and results in an implicit model for ρ_i.

A log likelihood function is derived by assuming that $z_i,\ i \in s^*$, are mutually independent given $\mathbf{x}_i$, $i \in s^*$. For an arbitrary parametric model for p_i, the resulting Elliott/ILR estimating function is

$$\hat{\mathbf{U}}_{\tilde{\pi}}^{E\text{-ILR}}(\boldsymbol{\alpha}) = \sum\nolimits_{k \in s_{\text{NP}}} \frac{1}{p_k(\boldsymbol{\alpha})} \frac{\tilde{\pi}_k}{\left[\tilde{\pi}_k + p_k(\boldsymbol{\alpha})\right]} \mathbf{g}_k(\boldsymbol{\alpha}) - \sum\nolimits_{k \in s_P} \frac{1}{\tilde{\pi}_k} \frac{\tilde{\pi}_k}{\left[\tilde{\pi}_k + p_k(\boldsymbol{\alpha})\right]} \mathbf{g}_k(\boldsymbol{\alpha}). \tag{4.1}$$

The estimating function (4.1) is ξmd unbiased. If $\tilde{\pi}_k$ is replaced by π_k in (4.1), the resulting estimating function is denoted by $\hat{\mathbf{U}}_{\pi}^{E\text{-ILR}}(\boldsymbol{\alpha})$. It is both md and ξmd unbiased provided that assumption (A2) holds. The estimating function (4.1) has a form similar to the optimal estimating function $\hat{\mathbf{U}}_{\tilde{\pi}}^{\text{opt}}(\boldsymbol{\alpha})$ given by (2.8) with π_k replaced by $\tilde{\pi}_k$. Both $\hat{\mathbf{U}}_{\tilde{\pi}}^{\text{opt}}(\boldsymbol{\alpha})$ and $\hat{\mathbf{U}}_{\tilde{\pi}}^{E\text{-ILR}}(\boldsymbol{\alpha})$ are expected to be roughly equivalent in general, except for scenarios where the sampling fraction in both samples is large and the overlap is not small. It is thus not surprising that the estimating function (4.1) performed better than the CLW and WVL estimating functions in the simulation study of Savitsky et al. (2022).

The two issues noted in Section 3 for the WVL/Pseudo-ILR method also apply for the Elliott/ILR method. The estimating function (4.1) does not have the CL property since it does not reduce to the population likelihood estimating function (2.1) when the probability sample is a census. Indeed, it reduces to the WVL/Pseudo-ILR estimating function (3.1).

Also, the assumption that $z_i,\ i \in s^*$, are mutually independent given $\mathbf{x}_i,\ i \in s^*$, is not valid since, using the setup of Savitsky et al. (2022) along with the random splitting of U^* described in Section 3,

$$\Pr\left(z_i = 1, z_j = 1 \middle| i, j \in s^*, \mathbf{x}_i, \mathbf{x}_j\right) = 0 \neq \frac{p_i}{(\tilde{\pi}_i + p_i)} \frac{p_j}{(\tilde{\pi}_j + p_j)},$$

for any pair of elements $i \in U_1$ and $j \in U_2$. For two different elements i and j in the same population, either U_1 or U_2, we also have

$$\Pr\left(z_i = 1, z_j = 1 \middle| i, j \in s^*, \mathbf{x}_i, \mathbf{x}_j\right) = \frac{p_i p_j}{\tilde{\pi}_i \tilde{\pi}_j + p_i p_j} \neq \frac{p_i}{(\tilde{\pi}_i + p_i)} \frac{p_j}{(\tilde{\pi}_j + p_j)},$$

when (A1) holds for elements $i \in U_{\text{NP}}$ as well as (A3) and (A5) for elements $i \in U_P$. Even under these assumptions, the mutual independence of $z_i,\ i \in s^*$, is not tenable unless many of the p_i's are small, and thereby the overlap is a negligible portion of the probability sample. This condition appears to be reasonably satisfied in the simulation study of Dr. Gershunskaya and Dr. Beresovsky and may explain the good performance of the AIC for the ILR method. In principle, an AIC based on an incorrect log likelihood function is not valid and may not be effective for variable selection.

5. A sample likelihood approach

5.1 Known overlap between both samples

A sample likelihood function (e.g., Pfeffermann, Krieger and Rinott, 1998) in the data integration scenario studied in our paper is a likelihood function based on observations from sample units $k \in s_{\text{NP}} \cup s_P$.

Let us first assume that we have access to $\mathbf{X}_s = \{\mathbf{x}_k; k \in s_{\mathrm{NP}} \cup s_P\}$ in addition to $\{\mathbf{x}_k; k \in s_{\mathrm{NP}}\}$. These two data sets do not need to be linked, but the overlap between the probability and non-probability samples needs to be known to create $\mathbf{X}_s$ from auxiliary data of the two samples. We thus assume that either $\{\delta_k, \mathbf{x}_k; k \in s_P\}$ or $\{I_k, \mathbf{x}_k; k \in s_{\mathrm{NP}}\}$ is known. This assumption will be relaxed in Section 5.2.

Under assumptions (A2) and (A4), it is easy to show that the probability of participation given $k \in s_{\mathrm{NP}} \cup s_P$ is

$$p_{s,k} = \Pr\left(\delta_k = 1 \middle| k \in s_{\mathrm{NP}} \cup s_P, \mathbf{x}_k\right) = \frac{\Pr\left(k \in s_{\mathrm{NP}} \middle| \mathbf{x}_k\right)}{\Pr\left(k \in s_{\mathrm{NP}} \cup s_P \middle| \mathbf{x}_k\right)} = \frac{p_k}{p_k + \tilde{\pi}_k - \tilde{\pi}_k p_k}. \tag{5.1}$$

This conditional participation probability reduces to $p_{s,k} \approx p_k / (p_k + \tilde{\pi}_k)$ when the overlap is negligible, which is the implicit assumption made by Elliott (2009). Note that our assumptions (A2) and (A4) do not necessarily imply a negligible overlap, in particular when the sampling fractions are large.

Under the independence assumptions (A1), (A3) and (A5), (I_k, δ_k), $k \in s_{\mathrm{NP}} \cup s_P$, are mutually independent given $\mathbf{x}_k$, $k \in s_{\mathrm{NP}} \cup s_P$. Assuming a parametric model for p_k, the sample likelihood function can be written as

$$L_s(\boldsymbol{\alpha}) = \prod_{k \in s_{\mathrm{NP}} \cup s_P} \left[p_{s,k}(\boldsymbol{\alpha})\right]^{\delta_k} \left[1 - p_{s,k}(\boldsymbol{\alpha})\right]^{(1-\delta_k)}, \tag{5.2}$$

where $p_{s,k}(\boldsymbol{\alpha})$ is given by (5.1) with $p_k = p_k(\boldsymbol{\alpha})$. If Poisson sampling is not used to select the probability sample, assumption (A3) does not hold. It remains to be verified if the sample likelihood function (5.2) would remain approximately valid for sampling designs used in practice beyond Poisson sampling. In the context of modelling probability sample data only, Pfeffermann, Krieger and Rinott (1998) showed the asymptotic independence of sample observations for common sampling designs provided that the population observations are independent. It is possible that a similar result would also hold when combining probability and non-probability sample data.

Using (5.2) and reorganising terms, we obtain the sample log likelihood function

$$l_s(\boldsymbol{\alpha}) = \sum_{k \in s_{\mathrm{NP}}} \log\left[p_{s,k}(\boldsymbol{\alpha})\right] + \sum_{k \in s_P} \log\left[1 - p_{s,k}(\boldsymbol{\alpha})\right] - \sum_{k \in s_{\mathrm{NP}} \cap s_P} \log\left[1 - p_{s,k}(\boldsymbol{\alpha})\right]. \tag{5.3}$$

Taking the derivative of $l_s(\boldsymbol{\alpha})$ with respect to $\boldsymbol{\alpha}$, we obtain, after straightforward algebra, the sample likelihood estimating function

$$\mathbf{U}_s(\boldsymbol{\alpha}) = \sum_{k \in s_{\mathrm{NP}}} \frac{1}{p_k(\boldsymbol{\alpha})} \frac{\tilde{\pi}_k p_{s,k}(\boldsymbol{\alpha})}{p_k(\boldsymbol{\alpha})} \mathbf{g}_k(\boldsymbol{\alpha}) - \sum_{k \in s_P} \frac{1}{\tilde{\pi}_k} \frac{\tilde{\pi}_k p_{s,k}(\boldsymbol{\alpha})}{p_k(\boldsymbol{\alpha})\left[1 - p_k(\boldsymbol{\alpha})\right]} \mathbf{g}_k(\boldsymbol{\alpha}) + \boldsymbol{\Psi}(\boldsymbol{\alpha}), \tag{5.4}$$

where

$$\boldsymbol{\Psi}(\boldsymbol{\alpha}) = \sum_{k \in s_{\mathrm{NP}} \cup s_P} I_k \delta_k \frac{p_{s,k}(\boldsymbol{\alpha}) \mathbf{g}_k(\boldsymbol{\alpha})}{p_k(\boldsymbol{\alpha})\left[1 - p_k(\boldsymbol{\alpha})\right]} = \sum_{k \in s_{\mathrm{NP}}} I_k \frac{p_{s,k}(\boldsymbol{\alpha}) \mathbf{g}_k(\boldsymbol{\alpha})}{p_k(\boldsymbol{\alpha})\left[1 - p_k(\boldsymbol{\alpha})\right]} = \sum_{k \in s_P} \delta_k \frac{p_{s,k}(\boldsymbol{\alpha}) \mathbf{g}_k(\boldsymbol{\alpha})}{p_k(\boldsymbol{\alpha})\left[1 - p_k(\boldsymbol{\alpha})\right]}. \tag{5.5}$$

The estimating function (5.4) satisfies the CL property and, under assumptions (A2) and (A4), is ξmd unbiased conditional on $\mathbf{X}_s$. From (5.5), we observe that the use of estimating function (5.4) requires knowing the overlap only in one of the two samples, i.e., observing either $\{I_k, \mathbf{x}_k; k \in s_{\text{NP}}\}$ or $\{\delta_k, \mathbf{x}_k; k \in s_P\}$ is sufficient. This information could be obtained via additional questions in the probability or non-probability survey, or via record linkage, with auxiliary variables being possible matching variables. For instance, if the vector $\mathbf{x}_k$ is distinct for each population unit $k \in U$ (e.g., there is at least one continuous auxiliary variable), it is possible to gain knowledge of $\{\delta_k, \mathbf{x}_k; k \in s_P\}$ and $\{I_k, \mathbf{x}_k; k \in s_{\text{NP}}\}$ by matching each unit of one sample with all the units of the other sample. That is, if the vector $\mathbf{x}_k$ for a unit $k \in s_P$ is identical to the vector $\mathbf{x}_l$ of a unit $l \in s_{\text{NP}}$, we then know that $\delta_k = 1$ (and $I_l = 1$). Otherwise, if there is no match with $\mathbf{x}_k$, then $\delta_k = 0$. This matching can be repeated for each unit $k \in s_P$ to identify the entire overlap $\{\delta_k, \mathbf{x}_k; k \in s_P\}$. A similar procedure can be used to identify $\{I_k, \mathbf{x}_k; k \in s_{\text{NP}}\}$. If sufficient information is available to implement the estimating function (5.4) then the classical AIC can be used for variable selection using the sample log likelihood function (5.3), i.e., $\text{AIC} = -2l_s(\hat{\boldsymbol{\alpha}}_s) + 2q$, where $\hat{\boldsymbol{\alpha}}_s$ is the solution of $\mathbf{U}_s(\boldsymbol{\alpha}) = \mathbf{0}$ and q is the number of model parameters. This is the ideal solution if the overlapping units can be accurately identified. Kim and Kwon (2024) independently proposed an unconditional propensity score model approach that appears to be very similar to our sample likelihood approach.

5.2 Unknown overlap between both samples

In practice, we may observe neither $\{\delta_k, \mathbf{x}_k; k \in s_P\}$ nor $\{I_k, \mathbf{x}_k; k \in s_{\text{NP}}\}$. One way to address this issue is by a direct application of the MIP. It consists of replacing the unobserved estimating function (5.4) with its expectation conditional on observed data, $\mathbf{X}_{\text{obs}} = \{\{\mathbf{x}_k; k \in s_{\text{NP}}\}, \{\mathbf{x}_k; k \in s_P\}\} \neq \mathbf{X}_s$. This leads to the estimating function

$$\begin{aligned} E_{\xi md}\left(\mathbf{U}_s(\boldsymbol{\alpha}) \middle| \mathbf{X}_{\text{obs}}\right) &= \sum_{k \in s_{\text{NP}}} \frac{1}{p_k(\boldsymbol{\alpha})} \frac{\tilde{\pi}_k p_{s,k}(\boldsymbol{\alpha}) \mathbf{g}_k(\boldsymbol{\alpha})}{p_k(\boldsymbol{\alpha})} \\ &\quad - \sum_{k \in s_P} \frac{1}{\tilde{\pi}_k} \frac{\tilde{\pi}_k p_{s,k}(\boldsymbol{\alpha}) \mathbf{g}_k(\boldsymbol{\alpha})}{p_k(\boldsymbol{\alpha})\left[1 - p_k(\boldsymbol{\alpha})\right]} + E_{\xi md}\left(\boldsymbol{\Psi}(\boldsymbol{\alpha}) \middle| \mathbf{X}_{\text{obs}}\right). \end{aligned} \tag{5.6}$$

Using the last term on the right-hand side of (5.5), the expectation $E_{\xi md}\left(\boldsymbol{\Psi}(\boldsymbol{\alpha}) \middle| \mathbf{X}_{\text{obs}}\right)$ in (5.6), which is the best predictor of $\boldsymbol{\Psi}(\boldsymbol{\alpha})$, can be written as

$$E_{\xi md}\left(\boldsymbol{\Psi}(\boldsymbol{\alpha}) \middle| \mathbf{X}_{\text{obs}}\right) = \sum_{k \in s_P} p_k^{\text{obs}} \frac{p_{s,k}(\boldsymbol{\alpha}) \mathbf{g}_k(\boldsymbol{\alpha})}{p_k(\boldsymbol{\alpha})\left[1 - p_k(\boldsymbol{\alpha})\right]},$$

where $p_k^{\text{obs}} = E_{\xi md}\left(\delta_k \middle| \mathbf{X}_{\text{obs}}\right)$, $k \in s_P$. Under our assumptions, it can be shown that $p_k^{\text{obs}} = E_{\xi md}\left(\delta_k \middle| n_k^{\text{NP}}\right) = n_k^{\text{NP}} / N_k$, where n_k^{NP}, $k \in s_P$, is the number of units $l \in s_{\text{NP}}$ for which $\mathbf{x}_l = \mathbf{x}_k$, and N_k, $k \in s_P$, is the number of units $l \in U$ for which $\mathbf{x}_l = \mathbf{x}_k$. The result follows by noting that n_k^{NP} obeys a binomial distribution with number of trials N_k and probability $p_k(\boldsymbol{\alpha})$. The application of the MIP in this context requires knowing N_k, $k \in s_P$. This information is typically unknown, but if we can assume that the population vectors $\mathbf{x}_k$ are all distinct (i.e., $N_k = 1$, $k \in U$), we can identify the entire overlap, as explained above, and thus $p_k^{\text{obs}} = \delta_k$, $k \in s_P$, and $E_{\xi md}\left(\boldsymbol{\Psi}(\boldsymbol{\alpha}) \middle| \mathbf{X}_{\text{obs}}\right) = \boldsymbol{\Psi}(\boldsymbol{\alpha})$. In other less trivial cases, N_k could be modelled using, for example, the Poisson distribution.

As a simpler alternative to modelling N_k, for situations where the overlap cannot be identified in any of the samples, we propose to replace $\boldsymbol{\Psi}(\boldsymbol{\alpha})$ in (5.4) by its BLU predictor. This may lead to somewhat reduced efficiency compared with the best predictor $E_{\xi md}\left(\boldsymbol{\Psi}(\boldsymbol{\alpha}) \,\middle|\, \mathbf{X}_{\text{obs}}\right)$, but at least it does not depend on unknown values N_k, $k \in s_P$, as shown below.

We consider the following linear unbiased predictor of $\boldsymbol{\Psi}(\boldsymbol{\alpha})$ that uses the available auxiliary data from both samples:

$$\hat{\boldsymbol{\Psi}}(\boldsymbol{\alpha}) = \sum_{k \in s_{\text{NP}}} \lambda_k \tilde{\pi}_k \frac{p_{s,k}(\boldsymbol{\alpha})\mathbf{g}_k(\boldsymbol{\alpha})}{p_k(\boldsymbol{\alpha})\left[1 - p_k(\boldsymbol{\alpha})\right]} + \sum_{k \in s_P} (1 - \lambda_k) p_k(\boldsymbol{\alpha}) \frac{p_{s,k}(\boldsymbol{\alpha})\mathbf{g}_k(\boldsymbol{\alpha})}{p_k(\boldsymbol{\alpha})\left[1 - p_k(\boldsymbol{\alpha})\right]}, \tag{5.7}$$

where λ_k, $k \in s_{\text{NP}} \cup s_P$, are constants. The estimator (5.7) is conditionally unbiased in the sense that $E_{\xi md}\left(\hat{\boldsymbol{\Psi}}(\boldsymbol{\alpha}) - \boldsymbol{\Psi}(\boldsymbol{\alpha}) \,\middle|\, \mathbf{X}_s\right) = \mathbf{0}$, provided that assumptions (A2) and (A4) hold. Replacing $\boldsymbol{\Psi}(\boldsymbol{\alpha})$ in (5.4) by the right-hand side of (5.7), we obtain the estimating function

$$\begin{aligned} \hat{\mathbf{U}}_s^{\lambda}(\boldsymbol{\alpha}) &= \sum_{k \in s_{\text{NP}}} \frac{1}{p_k(\boldsymbol{\alpha})} \left(1 + \lambda_k \frac{p_k(\boldsymbol{\alpha})}{1 - p_k(\boldsymbol{\alpha})}\right) \frac{\tilde{\pi}_k p_{s,k}(\boldsymbol{\alpha})\mathbf{g}_k(\boldsymbol{\alpha})}{p_k(\boldsymbol{\alpha})} \\ &\quad - \sum_{k \in s_P} \frac{1}{\tilde{\pi}_k} \left(1 + \lambda_k \frac{p_k(\boldsymbol{\alpha})}{1 - p_k(\boldsymbol{\alpha})}\right) \frac{\tilde{\pi}_k p_{s,k}(\boldsymbol{\alpha})\mathbf{g}_k(\boldsymbol{\alpha})}{p_k(\boldsymbol{\alpha})}. \end{aligned} \tag{5.8}$$

It is ξmd unbiased conditional on $\mathbf{X}_s$.

The BLU predictor of $\boldsymbol{\Psi}(\boldsymbol{\alpha})$ is obtained by determining λ_k, $k \in s_{\text{NP}} \cup s_P$, that minimize the prediction variance $\text{var}_{\xi md}\left(\mathbf{c}'\hat{\boldsymbol{\Psi}}(\boldsymbol{\alpha}) - \mathbf{c}'\boldsymbol{\Psi}(\boldsymbol{\alpha}) \,\middle|\, \mathbf{X}_s\right)$. Under our three independence assumptions, this prediction variance is minimized when $\text{var}_{\xi md}\left(\lambda_k \tilde{\pi}_k \delta_k + (1 - \lambda_k)\, p_k(\boldsymbol{\alpha})\, I_k - I_k \delta_k \,\middle|\, k \in s_{\text{NP}} \cup s_P, \mathbf{x}_k\right)$ is minimized for each $k \in s_{\text{NP}} \cup s_P$. The constant λ_k is thus determined so that $\lambda_k \tilde{\pi}_k \delta_k + (1 - \lambda_k) p_k(\boldsymbol{\alpha}) I_k$ predicts as accurately as possible $I_k \delta_k$, i.e., whether unit k is in the intersection of the two samples or not. Adding assumptions (A2) and (A4), it can be shown, after straightforward algebra, that the value of λ_k that minimizes the prediction variance is $\lambda_k = \lambda_{\tilde{\pi},k}^{\text{opt}}(\boldsymbol{\alpha}) = 1 - \gamma_{\tilde{\pi},k}^{\text{opt}}(\boldsymbol{\alpha})$, $k \in s_{\text{NP}} \cup s_P$, where $\gamma_{\tilde{\pi},k}^{\text{opt}}(\boldsymbol{\alpha})$ is given by (2.7) after replacing π_k by $\tilde{\pi}_k$. Using $\lambda_k = \lambda_{\tilde{\pi},k}^{\text{opt}}(\boldsymbol{\alpha})$ in (5.8), it turns out that the estimating function (5.8) reduces exactly to the optimal estimating function $\hat{\mathbf{U}}_{\tilde{\pi}}^{\text{opt}}(\boldsymbol{\alpha})$, which is given by (2.8) with π_k replaced by $\tilde{\pi}_k$. It is thus interesting to see that using BLU estimation theory under a population likelihood approach (see Section 2) is equivalent to using BLU prediction theory under a sample likelihood approach.

If the selection probability π_k is observed for all the probability and non-probability sample units, it can and should be considered as a potential auxiliary variable to be included in $\mathbf{x}_k$. If π_k is included in $\mathbf{x}_k$, $\tilde{\pi}_k = \pi_k$ and thus using π_k or $\tilde{\pi}_k$ in (2.8) does not make any difference. If π_k is not included in $\mathbf{x}_k$, because it does not appear to explain δ_k after conditioning on $\mathbf{x}_k$, then the above theory still remains valid and $\tilde{\pi}_k$ can be used in (2.8) as if π_k were unknown. It would also be possible to condition on both π_k and $\mathbf{x}_k$, which would result in replacing $\tilde{\pi}_k$ by π_k in the above developments. The optimal estimating function (2.8) would be md unbiased conditional on $\mathbf{X}_s$ (and unconditionally) provided that assumption (A2) holds.

5.3 Variable selection

The estimating function (2.8) is not the derivative of a sample log likelihood function, and thus the classical AIC is not applicable. Further research is needed on variable selection when $\hat{\mathbf{U}}_{\pi}^{\text{opt}}(\boldsymbol{\alpha})$ or $\hat{\mathbf{U}}_{\tilde{\pi}}^{\text{opt}}(\boldsymbol{\alpha})$ is used to estimate the participation probabilities. However, if many of the p_k's are small, the overlap is a negligible portion of the probability sample and the sample likelihood estimating function (5.4) becomes approximately equivalent to the estimating function $\hat{\mathbf{U}}_{\tilde{\pi}}^{\text{opt}}(\boldsymbol{\alpha})$. As a result, the sample log likelihood function (5.3), ignoring the negligible intersection term, can be used along with $\hat{\mathbf{U}}_{\tilde{\pi}}^{\text{opt}}(\boldsymbol{\alpha})$ to compute the classical AIC and select relevant auxiliary variables. It appears to be similar to the AIC Dr. Gershunskaya and Dr. Beresovsky used in their simulation study for the ILR method, except for the use of the estimating function $\hat{\mathbf{U}}_{\tilde{\pi}}^{E\text{-ILR}}(\boldsymbol{\alpha})$ given in (4.1). Both $\hat{\mathbf{U}}_{\tilde{\pi}}^{E\text{-ILR}}(\boldsymbol{\alpha})$ and $\hat{\mathbf{U}}_{\tilde{\pi}}^{\text{opt}}(\boldsymbol{\alpha})$ should be similar when the overlap is negligible. It is reassuring to see that their AIC performed well in their simulation study. We expect the performance to deteriorate as the non-probability sample size increases and the overlap becomes non-negligible.

6. A unified estimating function

Let us continue with the realistic scenario where neither $\{\delta_k, \mathbf{x}_k; k \in s_P\}$ nor $\{I_k, \mathbf{x}_k; k \in s_{\text{NP}}\}$ is known. In that scenario, we have described several methods in previous sections that led to different estimating functions. Assuming $\tilde{\pi}_k$ is used rather than π_k, they are all special cases of the general estimating function

$$\hat{\mathbf{U}}_{\tilde{\pi}}^{h}(\boldsymbol{\alpha}) = \sum\nolimits_{k \in s_{\text{NP}}} \frac{1}{p_k(\boldsymbol{\alpha})} h\left[\tilde{\pi}_k, p_k(\boldsymbol{\alpha})\right] \mathbf{g}_k(\boldsymbol{\alpha}) - \sum\nolimits_{k \in s_P} \tilde{w}_k h\left[\tilde{\pi}_k, p_k(\boldsymbol{\alpha})\right] \mathbf{g}_k(\boldsymbol{\alpha}), \tag{6.1}$$

where $h\left[\tilde{\pi}_k, p_k(\boldsymbol{\alpha})\right]$ is a function that depends on the method. Table 6.1 provides the expression of $h\left[\tilde{\pi}_k, p_k(\boldsymbol{\alpha})\right]$ for the methods described in previous sections.

Table 6.1
Expression of $h\left[\tilde{\pi}_k, p_k(\boldsymbol{\alpha})\right]$ for different methods.

Method	Estimating function	$h\left[\tilde{\pi}_k, p_k(\boldsymbol{\alpha})\right]$
CLW	$\hat{\mathbf{U}}_{\tilde{\pi}}(\boldsymbol{\alpha})$: (2.3)*	$\left[1 - p_k(\boldsymbol{\alpha})\right]^{-1}$
WVL/Pseudo-ILR	$\hat{\mathbf{U}}_{\tilde{\pi}}^{\text{WVL-PILR}}(\boldsymbol{\alpha})$: (3.1)*	$\left[1 + p_k(\boldsymbol{\alpha})\right]^{-1}$
Elliott/ILR	$\hat{\mathbf{U}}_{\tilde{\pi}}^{E\text{-ILR}}(\boldsymbol{\alpha})$: (4.1)	$\tilde{\pi}_k\left[\tilde{\pi}_k + p_k(\boldsymbol{\alpha})\right]^{-1}$
BLU estimation/prediction	$\hat{\mathbf{U}}_{\tilde{\pi}}^{\text{opt}}(\boldsymbol{\alpha})$: (2.8)**	$\tilde{\pi}_k\left[\tilde{\pi}_k + p_k(\boldsymbol{\alpha}) - 2\tilde{\pi}_k p_k(\boldsymbol{\alpha})\right]^{-1}$

* w_k is replaced with $\tilde{w}_k$ in (2.3) and (3.1).
** π_k is replaced with $\tilde{\pi}_k$ in (2.8).

Some authors (e.g., Beaumont, 2020; Chen, Li and Wu, 2020; and Rao, 2021) considered the calibration estimating function

$$\hat{\mathbf{U}}_{\pi}^{\text{cal}}(\boldsymbol{\alpha}) = \sum\nolimits_{k \in s_{\text{NP}}} \frac{1}{p_k(\boldsymbol{\alpha})} \mathbf{x}_k - \sum\nolimits_{k \in s_P} w_k \mathbf{x}_k.$$

Its smoothed version $\hat{\mathbf{U}}_{\tilde{\pi}}^{\text{cal}}(\boldsymbol{\alpha})$, obtained by replacing w_k with $\tilde{w}_k$ in the above equation, is also a special case of (6.1). For example, if a logistic model for $p_k(\boldsymbol{\alpha})$ is used, the estimating function (6.1) reduces to $\hat{\mathbf{U}}_{\tilde{\pi}}^{\text{cal}}(\boldsymbol{\alpha})$ when $h\left[\tilde{\pi}_k, p_k(\boldsymbol{\alpha})\right] = \left\{p_k(\boldsymbol{\alpha})\left[1 - p_k(\boldsymbol{\alpha})\right]\right\}^{-1}$. The calibration estimating function does not have the CL property but has an implicit double robustness property when a linear model between the survey variables and auxiliary variables holds. It could also be easily generalized to the scenario where different auxiliary variables are available in different probability samples as long as all the auxiliary variables are observed in the non-probability sample. The calibration estimating function $\hat{\mathbf{U}}_{\pi}^{\text{cal}}(\boldsymbol{\alpha})$ is the special case of (2.5) with $\gamma_k = \gamma_k^{\text{CAL}}(\boldsymbol{\alpha}) = -\left[1 - p_k(\boldsymbol{\alpha})\right] / p_k(\boldsymbol{\alpha}) < 0.$ It is thus expected to be inefficient for the estimation of $p_k(\boldsymbol{\alpha})$.

The estimating function (6.1) is ξmd unbiased, both unconditional and conditional on $\mathbf{X}_s$. The probability $\tilde{\pi}_k$ in (6.1) can also be replaced by π_k if it is available in the non-probability sample. The estimating function (6.1) remains (conditionally) ξmd unbiased provided that assumption (A2) holds (e.g., if π_k is included in $\mathbf{x}_k$). If π_k, $k \in U$, are treated as fixed, (6.1) is also (conditionally) md unbiased under assumption (A2). A hybrid estimating function that does not require the availability of π_k in the non-probability sample is

$$\hat{\mathbf{U}}_{\pi,\tilde{\pi}}^{h}(\boldsymbol{\alpha}) = \sum\nolimits_{k \in s_{\text{NP}}} \frac{1}{p_k(\boldsymbol{\alpha})} h\left[\tilde{\pi}_k, p_k(\boldsymbol{\alpha})\right] \mathbf{g}_k(\boldsymbol{\alpha}) - \sum\nolimits_{k \in s_P} w_k h\left[\tilde{\pi}_k, p_k(\boldsymbol{\alpha})\right] \mathbf{g}_k(\boldsymbol{\alpha}). \tag{6.2}$$

It is (conditionally) ξmd unbiased without requiring the validity of a model for π_k, but may be less efficient than (6.1) due to the variability of the probability survey weights w_k.

In practice, whether (6.1) or (6.2) is used, $\tilde{\pi}_k$ is unknown and must be estimated. As pointed out in Section 2, the probability sample can be used to estimate $\tilde{w}_k = E_{\xi}\left(w_k \middle| k \in s_P, \mathbf{x}_k\right)$ by $\hat{\tilde{w}}_k$, perhaps using nonparametric methods, such as machine learning methods. Using the relationship (2.9), $\tilde{\pi}_k$ is estimated by $\hat{\tilde{\pi}}_k = 1 \big/ \hat{\tilde{w}}_k$. Note that $\tilde{\pi}_k = E_{\xi}\left(\pi_k \middle| \mathbf{x}_k\right)$ cannot be estimated by modelling $E_{\xi}\left(\pi_k \middle| k \in s_P, \mathbf{x}_k\right)$ and ignoring the probability sampling design, as is sometimes suggested in the literature (e.g., Elliott, 2009; Elliott and Valliant, 2017). This is because the probability sampling design is (strongly) informative with respect to the distribution of π_k given $\mathbf{x}_k$.

Let us now consider the homogeneous group model for which the auxiliary variables partition the population into G groups and $p_k(\boldsymbol{\alpha}) = p_g$ for a unit k in group g. The smoothed weight for a unit k in group g is $\tilde{w}_g = E_{\xi}\left(w_k \middle| k \in s_{P,g}\right)$, where $s_{P,g}$ is the set of probability sample units that fall in group g. It can simply be estimated by the average of the weights in group g, i.e., $\hat{\tilde{w}}_g = \hat{N}_g \big/ n_g^P$, where $\hat{N}_g = \sum\nolimits_{k \in s_{P,g}} w_k$ and n_g^P is the probability sample size in group g. For a unit k in group g, $\hat{\tilde{\pi}}_k = \hat{\tilde{\pi}}_g = n_g^P \big/ \hat{N}_g$. Replacing $\tilde{\pi}_k$ by $\hat{\tilde{\pi}}_k$ in (6.1) or (6.2) and solving the estimating equations (either $\hat{\mathbf{U}}_{\tilde{\pi}}^{h}(\boldsymbol{\alpha}) = \mathbf{0}$ or $\hat{\mathbf{U}}_{\pi,\tilde{\pi}}^{h}(\boldsymbol{\alpha}) = \mathbf{0}$) for any choice of $h\left[\tilde{\pi}_k, p_k(\boldsymbol{\alpha})\right]$ yield $\hat{p}_g = n_g^{\text{NP}} \big/ \hat{N}_g$, the estimate of p_g, where n_g^{NP} is the non-probability sample size in group g. Instead, if $\tilde{\pi}_k$ is replaced by π_k in (6.1) or (6.2) and $h\left[\pi_k, p_k(\boldsymbol{\alpha})\right]$ depends on both π_k

and $p_k(\boldsymbol{\alpha})$, the resulting estimated participation probability in group g is not $\hat{p}_g = n_g^{\mathrm{NP}} \big/ \hat{N}_g$ anymore and does not have a closed form.

Note that $\hat{p}_g = n_g^{\mathrm{NP}} \big/ \hat{N}_g$ can be larger than 1. It is more likely to happen for large non-probability samples and small probability samples. For a general vector $\mathbf{x}_k$, this may suggest that a solution may exist less frequently with the logistic model ($p_k(\hat{\boldsymbol{\alpha}})$ bounded by 1) than with the exponential model ($p_k(\hat{\boldsymbol{\alpha}})$ unbounded). However, the exponential model may not be accurate for large non-probability samples.

For the homogeneous group model, solving the sample likelihood estimating equation $\mathbf{U}_s(\boldsymbol{\alpha}) = \mathbf{0}$, where $\mathbf{U}_s(\boldsymbol{\alpha})$ is given in (5.4), yields

$$\hat{p}_g^{\mathrm{SL}} = \frac{\hat{p}_g}{\hat{p}_g + \left(1 - \frac{n_g^I}{n_g^P}\right)}$$

as the estimate of p_g, where n_g^I is the number of units in the intersection $s_{\mathrm{NP}} \cap s_P$ that fall in group g and $\hat{p}_g = n_g^{\mathrm{NP}} \big/ \hat{N}_g$. As expected, $\hat{p}_g^{\mathrm{SL}}$ is close to $\hat{p}_g$ when $\hat{p}_g$ and the overlap rate $n_g^I \big/ n_g^P$ are small. The sample likelihood estimate $\hat{p}_g^{\mathrm{SL}}$ cannot be greater than 1 unlike $\hat{p}_g$. This is a desirable property of the sample likelihood estimating function (5.4), which results from exploiting information on the overlap between both samples.

7. Concluding remarks

In previous sections, we described three likelihood approaches for the estimation of participation probabilities and selection of relevant auxiliary variables that are valid regardless of the size of the probability and non-probability samples as well as the size of the overlap between both samples: the population and pseudo likelihood approaches, described in Section 2, and the sample likelihood approach, described in Section 5. If the probability sample is a census, the population likelihood approach is the most efficient and should be the preferred choice. If the probability sample is not a census, but the overlap is known in at least one of the samples, the sample likelihood approach should be preferred over the pseudo likelihood approach for efficiency considerations. If the overlap is unknown, the pseudo likelihood approach of Chen, Li and Wu (2020) can be used both for the estimation of participation probabilities and computation of an AIC for variable selection. However, the CLW estimating function (2.3) may not be efficient, especially when the non-probability sample is larger than the probability sample, because it does not fully leverage the available auxiliary data. Our optimal estimating function (2.8), $\hat{\mathbf{U}}_\pi^{\mathrm{opt}}(\boldsymbol{\alpha})$, or its smoothed version $\hat{\mathbf{U}}_{\tilde{\pi}}^{\mathrm{opt}}(\boldsymbol{\alpha})$, is expected to be more efficient than existing alternatives, although it remains to be demonstrated in an empirical study. Variable selection in the case of unknown overlap requires further research when $\hat{\mathbf{U}}_\pi^{\mathrm{opt}}(\boldsymbol{\alpha})$ or $\hat{\mathbf{U}}_{\tilde{\pi}}^{\mathrm{opt}}(\boldsymbol{\alpha})$ is used, except for the case where many of the participation probabilities are small and the overlap can be neglected. In that case, the sample log likelihood function (5.3), ignoring the overlap term, along with $\hat{\mathbf{U}}_{\tilde{\pi}}^{\mathrm{opt}}(\boldsymbol{\alpha})$ can be used to compute the classical AIC.

In practice, estimated participation probabilities from a parametric model are rarely used directly to compute estimates of finite population parameters. Groups homogeneous with respect to these estimated probabilities are often created to protect against model misspecifications and extreme inverse probability weights. It is possible that the choice of an estimating function does not have a major impact on the estimates of finite population parameters if homogeneous groups are used before computing those estimates. Nevertheless, it seems reasonable to choose the most efficient estimating function for the estimation of participation probabilities before creating homogeneous groups.

Nonparametric estimation of participation probabilities using, for example, machine learning methods could be useful to protect against possible model misspecifications. Existing machine learning methods can be directly used to model p_k if $\left\{\delta_k, \mathbf{x}_k; k \in U\right\}$ is known. Otherwise, if $\left\{\delta_k, \mathbf{x}_k; k \in s_{\mathrm{NP}} \cup s_P\right\}$ is known, the conditional probability $p_{s,k}$ can be modelled using existing machine learning methods and p_k can then be estimated using relationship (5.1). The most difficult case is when the overlap is unknown. In our main paper, we proposed nppCART as a means of creating homogeneous groups and obtaining protection against model misspecifications. Our procedure is inspired from the pseudo likelihood approach of Chen, Li and Wu (2020) and does not require a negligible overlap between both samples. An alternative would be to consider machine learning methods along with the Elliott/ILR method, as suggested in Elliott and Valliant (2017) and Elliott (2022). The idea would consist of modelling $\rho_i = \Pr\left(z_i = 1 \middle| i \in s^*, \mathbf{x}_i\right)$ using a machine learning method, ignoring the overlap and thus the lack of independence between observations. Then, p_i would be estimated for the non-probability sample units using the relationship $\rho_i = p_i / (\tilde{\pi}_i + p_i)$, whose validity was shown in Section 4 using the setup of Savitsky et al. (2022). If most of the participation probabilities are small, the overlap is a negligible portion of the probability sample and can thus be ignored. Therefore, this approach would be equivalent to our suggestion above of modelling $p_{s,k}$ using a machine learning method and then using relationship (5.1) to estimate p_k. It remains to be evaluated how that machine learning version of the Elliott/ILR method would perform when the overlap is not negligible.

In our main paper and in this rejoinder, we have focussed on the estimation of the participation probability p_k for non-probability sample units. Once the estimates $\hat{p}_k$, $k \in s_{\mathrm{NP}}$, are computed, they can be used to estimate finite population parameters, such as population totals or means. The basic inverse probability weighted estimator of finite population parameters consists of weighting non-probability sample units by $1/\hat{p}_k$. Of course, there are estimators that more efficiently use $\hat{p}_k$, for instance, by taking advantage of a model for the survey variables to achieve a double robustness property (e.g., Chen, Li and Wu, 2020; Chambers, Ranjbar, Salvati and Pacini, 2022). The simplest, but common, example is when the inverse probability weights $1/\hat{p}_k$ are calibrated on known or estimated population totals of auxiliary variables. The resulting estimator of population totals is doubly robust in the sense that it is valid under either the participation model or a linear model between survey variables and auxiliary variables.

Our point of view is that survey statisticians should start with the most efficient estimates of p_k, $k \in s_{\mathrm{NP}}$, possible before using them for the estimation of finite population parameters. This is exactly the same point of view many survey statisticians take for the estimation of finite population parameters using data from a probability sample; they start with their best estimate of the probability of selection in the sample, π_k, and then use it to derive efficient estimators of finite population parameters (e.g., using calibration techniques). It just happens that the probability π_k is usually known for probability samples and does not require to be estimated.

In this final remark, we would like to take this opportunity to sincerely thank Prof. Partha Lahiri, the guest editor of this special issue, for all his efforts in organizing such a nice collection of papers, which were presented at the 2022 Morris Hansen lecture event, along with their discussion.

References

Beaumont, J.-F. (2008). A new approach to weighting and inference in sample surveys. *Biometrika*, 95, 539-553.

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.

Beresovsky, V. (2019). On application of a response propensity model to estimation from web samples. In ResearchGate.

Chambers, R.L. (2023). The missing information principle – A paradigm for analysis of messy sample survey data. *Survey Methodology*, 49, 2, 219-256. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2023002/article/00018-eng.pdf.

Chambers, R., Ranjbar, S., Salvati, N. and Pacini, B. (2022). Weighting, informativeness and causal inference, with an application to rainfall enhancement. *Journal of the Royal Statistical Society, Series A (Statistics in Society)*, 185, 1584-1612.

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

Elliott, M.R. (2009). Combining data from probability and non-probability samples using pseudo-weights. *Survey Practice*, 2, 813-845.

Elliott, M.R. (2022). Comments on "Statistical inference with non-probability survey samples". *Survey Methodology*, 48, 2, 319-329. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2022002/article/00004-eng.pdf.

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

Gershunskaya, J., and Lahiri, P. (2023). Discussion of "Probability vs. nonprobability sampling: From the birth of survey sampling to the present day", by Graham Kalton. *Statistics in Transition New Series*, 24, 3, 31-37.

Kim, J.K., and Kwon, Y. (2024). Comment on "Exchangeability assumption in propensity-score based adjustment methods for population mean estimation using non-probability samples". *Survey Methodology*, 50, 1, 57-63. Paper available at http://www.statcan.gc.ca/pub/12-001-x/2024001/article/00007-eng.pdf.

Lumley, T., and Scott, A. (2015). AIC and BIC for modeling with complex survey data. *Journal of Survey Statistics and Methodology*, 3, 1-18.

Orchard, T., and Woodbury, M.A. (1972). A missing information principle: Theory and application. *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability*, 1, 697-715.

Pfeffermann, D., Krieger, A.M. and Rinott, Y. (1998). Parametric distributions of complex survey data under informative probability sampling. *Statistica Sinica*, 8, 1087-1114.

Pfeffermann, D., and Sverchkov, M. (1999). Parametric and semi-parametric estimation of regression models fitted to survey data. *Sankhyā: The Indian Journal of Statistics, Series B*, 61, 166-186.

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

Savitsky, T.D., Williams, M.R., Gershunskaya, J., Beresovsky, V. and Johnson, N.G. (2022). Methods for combining probability and nonprobability samples under unknown overlaps. https://doi.org/10.48550/arXiv.2208.14541.

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.

JOURNAL OF OFFICIAL STATISTICS

An International Review Published by Statistics Sweden

JOS is a scholarly quarterly that specializes in statistical methodology and applications. Survey methodology and other issues pertinent to the production of statistics at national offices and other statistical organizations are emphasized. All manuscripts are rigorously reviewed by independent referees and members of the Editorial Board.

Contents
Volume 39, No. 4, December 2023

All inquires about submissions and subscriptions should be directed to jos@scb.se

JOURNAL OF OFFICIAL STATISTICS

An International Review Published by Statistics Sweden

JOS is a scholarly quarterly that specializes in statistical methodology and applications. Survey methodology and other issues pertinent to the production of statistics at national offices and other statistical organizations are emphasized. All manuscripts are rigorously reviewed by independent referees and members of the Editorial Board.

Contents
Volume 40, No. 1, March 2024

Letter to Editor

Articles

All inquires about submissions and subscriptions should be directed to jos@scb.se

The Canadian Journal of Statistics **La revue canadienne de statistique**

CONTENTS TABLE DES MATIÈRES

Volume 51, No. 3, September/septembre 2023

Special Issue in Honour of Nancy Reid/Numéro spécial en l’honneur de Nancy Reid

Introduction

Discussion

Review Article

Research Articles

The Canadian Journal of Statistics **La revue canadienne de statistique**

CONTENTS TABLE DES MATIÈRES

Volume 51, No. 4, December/décembre 2023

Research Article

GUIDELINES FOR MANUSCRIPTS

Authors are invited to submit their articles **through the *Survey Methodology* hub on the ScholarOne Manuscripts website** (https://mc04.manuscriptcentral.com/surveymeth). Before submitting the article, please examine a recent issue of *Survey Methodology* as a guide and note particularly the points below. Articles must be submitted in Word or Latex, preferably in Word with MathType for the mathematical expressions. A pdf version is also required for formulas and figures.

1. Layout

1.1 Documents should be typed entirely double spaced with margins of at least 1½ inches on all sides.

1.2 The documents should be divided into numbered sections with suitable verbal titles.

1.3 The name (fully spelled out) and address of each author should be given as a footnote on the first page of the manuscript.

1.4 Acknowledgements should appear at the end of the text.

1.5 Any appendix should be placed after the acknowledgements but before the list of references.

2. Abstract and Introduction

2.1 The manuscript should begin with an abstract consisting of one paragraph followed by three to six key words. Avoid mathematical expressions in the abstract.

2.2 The last paragraph of the introduction should contain a brief description of each section.

3. Style

3.1 Avoid footnotes and abbreviations.

3.2 Limit the use of acronyms. If an acronym is used, it must be defined the first time it occurs in the paper.

3.3 Mathematical symbols will be italicized unless specified otherwise except for functional symbols such as “$\exp(\cdot)$” and “$\log(\cdot)$”, etc.

3.4 Short formulae should be left in the text but everything in the text should fit in single spacing. Long and important equations should be separated from the text and numbered with arabic numerals on the right if they are to be referred to later. Use a two-level numbering system based on the section of the paper. For example, equation (4.2) is the second important equation in Section 4.

3.5 Bold fonts should normally be used to distinguish vectors and matrices from scalars.

4. Figures and Tables

4.1 All figures and tables should be numbered with arabic numerals, with titles that are as self explanatory as possible, at the top of tables or figures. Use a two-level numbering system based on the section of the paper. For example, Table 3.1 is the first table in Section 3.

4.2 A detailed textual description of figures may be required for accessibility purposes if the message conveyed by the image is not sufficiently explained in the text.

5. References

5.1 References in the text should be cited with authors’ names and the date of publication. If part of a reference is cited, indicate after the reference, e.g., Cochran (1977, page 164).

5.2 The first time a reference is cited in the text, the name of all authors must be written. For subsequent occurrences, the names of all authors can again be written. However, if the reference contains three or more authors, the names of the second and subsequent authors can be replaced with “et al.”.

5.3 The list of references at the end of the manuscript should be arranged alphabetically and for the same author chronologically. Distinguish publications of the same author in the same year by attaching a, b, c to the year of publication. Journal titles should not be abbreviated. Follow the same format used in recent issues.

6. Short Notes

6.1 Documents submitted for the short notes section must have a maximum of 3,000 words, including tables, figures and references.