UNIVERSITY OF CALIFORNIA,
IRVINE

On Generalizable Inference and Prediction for Biased Samples

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Statistics

by

Olivia Marie Bernstein Morgan

Dissertation Committee:
Professor Daniel L. Gillen, Chair
Doctor Brian G. Vegetabile
Professor Zhaoxia Yu

2022

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I would like to thank my advisor, Daniel Gillen, for all of your support and advice during graduate school. I have learned so much from you about how to answer scientific questions and the ethics for doing research with human volunteers. I really value your insight. Thank you for spending hours talking through career decisions and encouraging me when I thought I could not solve a problem.

I would also like to thank Brian Vegetabile and Joshua Grill for their help in advising me. Brian, I have appreciated your help troubleshooting errors and advice about career and school over the years. Josh, I learned so much about Alzheimer's Disease, careers in research, and scientific writing from you. Thank you to Zhaoxia Yu for being on my dissertation committee and for being an excellent teacher.

Thank you the members of the Grillen lab and all of the students in the Statistics Department at UCI, especially those in my cohort. I could not have done it without your support and encouragement. I want to thank Michelle Nuño and Mary Ryan for letting me follow in your footsteps and telling me things would work out. I also want to thank Hannah Baumann and Jessica Pazienza for listening to me talk through my research and for giving suggestions.

I would like to thank my friends and family for their support. My parents always encouraged me to keep going. Thank you to my husband, Grant Morgan, for always listening and supporting me. I am so thankful for you.

The text of Chapter 3 of this dissertation is a modified reprint of the material as it appears in Health Services and Outcomes Research Methodology [15], used with permission from Springer Nature. The co-authors listed in this publication directed and supervised research which forms the basis for the dissertation.

The research in this dissertation was funded by the National Science Foundation Graduate Research Fellowship under Grant No. DGE-183928 and the ARCS foundation. It was also supported by the National Institute on Aging of the National Institute of Health awards 1RF1AG059407 and 1R01AG077628-01. I would like to thank the participants of the Consent-to-Contact (C2C) registry and the Alzheimer's Disease Research Centers (ADRC). The C2C registry was made possible through a philanthropic gift from HCP, Inc and grants P30 AG066519 and UL1 TR000153. The National Alzheimer's Coordinating Center (NACC) database is funded by NIA/NIH Grant U24 AG072122. NACC data are contributed by the NIA-funded ADRCs. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

VITA

Olivia Marie Bernstein Morgan

EDUCATION

Doctor of Philosophy in Statistics **2022**
University of California, Irvine *Irvine, California*

Master of Science in Statistics **2019**
University of California, Irvine *Irvine, California*

Bachelor of Science in Chemistry and Mathematics **2017**
Biola University *La Mirada, California*

RESEARCH EXPERIENCE

Graduate Research Assistant **2018–2022**
University of California, Irvine *Irvine, California*

Summer Undergraduate Research Fellow **2016**
University of Georgia, Athens *Athens, Georgia*

Undergraduate Research Assistant **2016**
Biola Univeristy *La Mirada, California*

TEACHING EXPERIENCE

Teaching Assistant **2017–2018**
University of California, Irvine *Irvine, California*

REFEREED JOURNAL PUBLICATIONS

Adjustment for Biased Sampling Using NHANES Derived Propensity Weights **2022**
Health Services and Outcomes Research Methodology

Anxiety and depressive symptoms and cortical amyloid-β burden in cognitively unimpaired older adults **2022**
The Journal of Prevention of Alzheimer's Disease

Recruitment and retention of participant and study partner dyads in two multinational Alzheimer's disease registration trials **2021**
Alzheimer's Research & Therapy

Education and message framing increase willingness to undergo research lumbar puncture: A randomized controlled trial **2021**
Frontiers in Geriatric Medicine

Reinterpreting the infrared spectrum of H + HCN: Methylene amidogen radical and its coproducts **2018**
The Journal of Chemical Physics

Fluorescent pseudorotaxanes of a quinodicarbocyanine dye with gamma cyclodextrin **2018**
Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy

SOFTWARE

`estweight` **R package** `https://github.com/oliviabern/estweight`
R package that estimates sampling weights for convenience samples using information from a representative sample.

AWARDS AND HONORS

Graduation Research Fellowship **2019-2022**
National Science Foundation

ARCS Foundation Scholar **2019-2021**
Achievement Rewards for College Scientists Foundation

Outstanding Graduate Teaching Assistant Award **2018**
Statistics Department, University of California, Irvine

ABSTRACT OF THE DISSERTATION

On Generalizable Inference and Prediction for Biased Samples

By

Olivia Marie Bernstein Morgan

Doctor of Philosophy in Statistics

University of California, Irvine, 2022

Professor Daniel L. Gillen, Chair

Most statistical methods assume that samples are representative of a target population of interest, but this assumption is commonly violated in biomedical applications with human volunteers. Study participants self-selection into a sample can cause it to be unrepresentative which in turn leads to sampling bias. When analyzing data from a biased sample, the sampling scheme must be accounted for to obtain inference and predictions that generalize to the target population. In this dissertation, we discuss an approach for addressing sampling bias by estimating sampling weights using an auxiliary data set. We then apply estimated sampling weights in the field of causal inference. We assess the impact on bias and variance of estimated causal effects when sampling weights are included or omitted when estimating propensity scores and propensity-adjusted causal effects. Lastly, we quantify the impact of sampling bias on estimates of the prediction error for a target population and compare estimation methods.

Chapter 1

Introduction

Most statistical methods assume samples are representative of some target population, but this assumption is often violated in practice. It can be difficult or unethical to collect a random sample from the target population. Medical research is conducted on samples of volunteers which tend to overrepresent non-Hispanic White and male participants but this does not represent the diversity of the United States population [87]. The sampling scheme needs to be accounted for when analyzing unrepresentative samples to estimate associations for some of interest. Sampling bias is the bias in estimated associations that arises when using a sample that is not representative of some population. Accounting for how the sample was collected will lead to more generalizable inference and predictions for the target population so everyone in the target population can equally benefit from medical research.

Sampling bias can have real world consequences. A recent paper by Obermeyer et al. (2019) [86] assessed the performance of a prediction model used to identify patients for high-risk care management. They found that the predicted risk of future care needs was less accurate for Black patients. More specifically, Black patients with more active chronic conditions were less likely to be identified for high-risk care management programs compared to White patients.

Failing to identify black patients for the program could lead to poorer health outcomes. Obermeyer et al. found that this happened because the model used healthcare spending as a proxy for health status, but Black patients had lower health care costs than White patients with the same number of active chronic conditions. Thus, the relationship between healthcare spending and health (as measured by the number of active chronic conditions) is differential for Black and White patients. When differential relationships like this are marginalized over, they can lead to bias in predictions if subpopulations, such as the Black patients in this example, are misrepresented in the training sample. Considering the sampling scheme and addressing it is vital for estimating predictions with the same error rate for all subpopulations.

In this dissertation, we develop methods for estimating sampling weights for biased samples and implementing them into analyses when the sampling probabilities were not prespecified by design. The objective of my research is to develop methods for valid estimates of associations and predictions that generalize to a target population when using biased samples while quantifying uncertainty.

We begin this dissertation in Chapter 2 with a background on different sampling schemes and on the use of survey samples and inverse probability weighting to account for biased sampling designs. Next, we provide a background on causal inference including a framework for estimating causal effects, the assumptions involved, and methods for isolating causal effects in the absence of randomization. Third, we give an overview of existing methods for estimating out-of-sample prediction error in simple random samples and biased samples. This chapter provides the necessary background for the methods developed and assessed throughout the remainder of the dissertation.

In Chapter 3, we compare methods for estimating sampling weights for biased samples using a representative sample for calibrating weights. We propose using the National Health and Nutrition Examination Survey (NHANES) [25] as a representative sample for biomedical applications because it is open access and collects medical measurements on participants. We provide a practical approach for utilizing NHANES to answer population questions via biased samples. We further propose methods to estimate the variance of parameter estimates that accounts for uncertainty that arises from the fact that sampling weights are estimated. Simulation studies explore the impact of sampling weight estimation on uncertainty in coefficient estimates in generalized linear models. We provide an R package to estimate coefficients for generalized linear models and corresponding variance estimates. We then apply these methods to obtain valid population-level estimates of racial and ethnic differences in willingness to be contacted about research opportunities using participants from the Consent-to-Contact (C2C) registry at the University of California, Irvine.

In Chapter 4, we discuss methods for incorporating estimated sampling weights into propensity-adjusted estimates of causal effects when using biased samples. The objective of this chapter is to estimate marginal treatment effects for a target population using a convenience sample when there is treatment effect heterogeneity among over- or underrepresented subpopulations. We conduct a simulation study to assess the utility of estimating sampling weights for convenience samples with an auxiliary dataset and implementing them into models for propensity score estimation, treatment effect estimation, or both. We also evaluate the ability to estimate sampling weights and treatment effects with a convenience sample when the full sampling scheme is not measured and a proxy is observed instead. Additionally, we derive an analytic variance estimator for estimated treatment effects when sampling weights are included in the propensity score and outcome models using simultaneous estimating equations approach that accounts for the uncertainty from estimating sampling weights and propensity scores and quantify its performance against the empirical variance. We provide an R package to estimate causal effects using propensity scores in a biased sample and corresponding variance estimates. We apply our approach to data from the National Alzheimer's Coordinating Center (NACC) Uniform Data Set [13] which is not a simple random sample of the United States population because most volunteers come from referrals. We use the NACC data and our method to estimate the effect of vitamin E supplementation on functional activities with estimated sampling weights derived using NHANES.

In Chapter 5, we assess methods for estimating the out-of-sample prediction error in a set target population when predictions models are trained and evaluated with biased samples that are too small for a hold out test set. We compare two classes of estimates: analytic and resampling estimates. Analytic methods estimate the prediction error using an analytic relationship between the training error and prediction error for biased samples. Resampling methods mimic estimating the prediction error with a separate test set by resampling the data and reweighting contributions to the loss function. To our knowledge, no one has compared the performance of these estimators in a biased sample so we conduct a simulation study to evaluate the merits of each method. Additionally, we compare their performance for different samples sizes and when sampling weights are known versus estimated. C2C researchers are planning to modify recruitment efforts to oversample neighborhoods with more socioeconomic disadvantage which will limit generalizability of prediction models fit in C2C if the sampling scheme is not accounted for. We compare estimates of the analytic and resampling estimates of the prediction error for models fit with a subsample from C2C drawn with sampling probabilities that are a function of socioeconomic disadvantage.

We conclude with a summary of our contributions and findings in Chapter 6. We then discuss future research directions to extend methodology for obtaining generalizable inference and predictions when using biased samples.

Chapter 2

Background

2.1 Survey Sampling

A main goal of statistics is to use a sample to learn about a population. Many statistical methods assume observations in the sample are representative of a set population of interest, or the target population. This assumption can be violated in many ways. For example, sampling units with equal probability requires a list of all units, known as a sampling frame, of the target population, but this may not exist [73]. Alternatively, even if the list of all units is available it may be too expensive or time consuming to sample all individuals with equal probability. Lumley (2010) [75] gives an example of the National Health and Nutrition Examination Survey (NHANES). NHANES samples approximately 5,000 people per year [25]. Sampling uniformly from the United States would be impractical because participants need to visit mobile examination sites across the country. It would be more efficient to sample multiple people in each geographic location. Survey samples are useful because they can increase statistical efficiency of estimates and efficiency in terms of saving time or resources [75].

2.1.1 Sample Types

Probability samples

Probability samples are samples where each unit's sampling probability is known and prespecified by design in the sampling frame and each unit is drawn randomly according to their sampling probability [73]. We assume there is a finite population we want to learn about and a sample that is a subset of a finite population. The most straight forward example of a probability sample is a simple random sample (SRS) where each subset of units of size n from the sampling frame has an equal probability of being the sample. A stratified random sample is one where the population is divided into disjoint subsets called strata and a probability sample is drawn from each strata. This ensures representation from each strata and can increase efficiency. A third type of probability sample is the cluster random sample where units are part of a grouping called a cluster. One example of a cluster is a city where the units are the citizens of that city. A probability sample of the clusters is drawn and then a probability sample is drawn within each sampled cluster.

Non-probability samples

Non-probability samples are ones where the researchers do not know the sampling probabilities for each unit in the sampling frame and did not randomly select units for the sample. A convenience sample is one where the units from the population who are most likely to respond are overrepresented [73]. This leads to a sample that is not representative of a desired target population at best and one where sampling probabilities of zero can exist for some populations at worst. For example, the National Alzheimer's Coordinating Center (NACC) Uniform Data Set recruits participants through their physician [13]. The NACC website clearly states that inference based on the NACC Data Set may not generalize to the population of the United States.

The objective of methods developed for analyzing survey samples is to estimate quantities for some finite population. The methods typically assume that the sampling probabilities are known and prespecified. As such the methods discussed in this section only apply to probability samples, but they can be extended to convenience samples by estimating sampling sampling probabilities relative to an auxiliary sample that is representative of the target population [2, 15, 26, 94, 40, 117, 88]. In practice, it is difficult to obtain a representative sample because it is expensive and may even be infeasible. There is need of a representative data set of the United States that collects demographic information and medical measurements to estimate sampling probabilities for biomedical studies done in a convenience sample. In Chapter 3, we discuss the utility of using NHANES as a representative sample for these types of studies.

2.1.2 Design-Based and Model-Based Frameworks

There are two common frameworks for analyzing samples: design-based and model-based inference [75]. Model-based inference assume data is generated from a probability model. For example, let y_i be the i-th observation in a sample, then y_i could follow a normal distribution with some mean μ and variance σ^2: $y_i \sim \text{Normal}(\mu, \sigma^2)$. This framework is assumed for many statistical methods. Survey sampling, however, often assumes a design-based framework. This framework assumes there is a finite population with N observations and the i-th observation y_i in the population is fixed for all i. Randomness comes from the sampling scheme and how units are sampled. Sometimes the distinctions between design-based and model-based frameworks are unclear and they are combined. An example of this is regression models for survey samples discussed in Section 2.1.4.

Assumptions for design-based inference

There are four assumptions required for design-based inference of population parameters[75]. Let π_i be the sampling probability for the i-th unit in the finite population and π_{ij} be the joint sampling probability for units i and j. Let $\mathcal{S}$ and $\mathcal{P}$ be the sets of individuals in the sample and the finite population. These sets have cardinality of n and N, respectively. The assumptions are:

1. Each unit in the population must have a non-zero sampling probability: $\pi_i > 0 \; \forall \; i \in \mathcal{P}$.
2. The sampling probability π_i must be known for each unit in the sample.
3. Each pair of units i and j in the sample must have a non-zero probability of both being sampled π_{ij}.
4. The joint sampling probability π_{ij} must be known for every pair in the sample.

If units are sampled independently then assumptions 1 and 2 imply 3 and 4. The joint sampling probabilities are usually calculated with statistical software for analyzing survey data based on the survey design[75].

2.1.3 Sampling Weights

Sampling weights are commonly used to account for the sampling scheme in design-based inference for finite populations [75]. Sampling weights are inversely proportional to sampling probabilities. The sampling weight, w_i for observation i is

$$w_i \propto \frac{1}{\pi_i}. \tag{2.1}$$

Sampling weights that are inversely proportional to the sampling probability are called inverse probability weights (IPW)[75]. Sampling weights can be scaled in several ways but it common to scale them so the sum of weights for the n observations in a sample sum to 1, n, or N. If the weights are scaled so they sum to the finite population size, N, then w_i is the number of units in the finite population the sampled observation i represents. Note that sampling weights can be estimated for convenience samples by first estimating the sampling probability as discussed in Section 2.1.1.

Horvitz-Thompson estimator

Survey weights allow us to estimate population quantities with a sample drawn with unequal sampling probabilities [75]. Consider the population total of variable x_i,

$$T = \sum_{i \in \mathcal{P}} x_i.$$

This can be estimated with the Horvitz-Thompson (HT) estimator

$$\widehat{T} = \sum_{i \in \mathcal{S}} N \times w_i \times x_i.$$

where $\sum_{i=1}^{n} w_i = 1$. This estimator is named for Horvitz and Thompson who formalized it and derived the variance estimate [59]. The HT estimator can be extended for other estimands such as the population mean.

Using a survey sample and accounting for the sampling scheme instead of using a SRS impacts the uncertainty of estimates [75]. The design effect is the ratio of the variance of the estimate from a survey sample divided by the variance of the estimate from a SRS is the design effect [64]. This quantity is useful to compute because it tells you the ratio of the sample size of a complex sampling design to the SRS needed to obtain the same variance estimate.

2.1.4 Regression Models for Survey Samples

Survey weights are also used in regression models to estimate associations for the target finite population using a survey sample [77]. Suppose we have a sample of n observations where x_i is a vector of covariates or predictors, y_i is a response variable, and $i = 1, \ldots, n$. Each observation has a sampling weight w_i. Suppose we are interested in fitting a generalized linear model (GLM) [85]. To estimate associations in a target population we can consider a hybrid of the design-based and model-based frameworks. We assume that our finite population is a subset of a larger super population, which is a subset of an even larger super population, and so forth until the populations are as large as necessary for asymptotic results[73].

Coefficient estimates in survey regression models

The remainder of this section is based on Lumley and Scott (2017) [77]. Suppose we have a super population which is a realization of a probability model $f(Y|X;\beta)$. The finite population is a realization from this super population. The sample is then drawn from the finite population and the only randomness in this stage is from the sampling. Let $g(\cdot)$ be a link function and β be a $p \times 1$ vector of parameters. Then

$$g(E[Y|X = x]) = g(\mu) = x^T\beta$$

and the variance of the response is a function of the mean μ where

$$\text{Var}(Y|X = x) = \sigma^2 V(\mu).$$

We can assume that $n \to \infty$ and $n/N \to c \in [0, 1)$. Let β_0 be the true paramater in the super population, $\tilde{\beta}_N$ be the estimate for β_0 that would be estimated using the full finite population, and $\widehat{\beta}_n$ be the maximum pseudo-likelihood estimate that is obtained using the survey sample when the sampling scheme is accounted for. Additionally, let $\tilde{\beta}_N \xrightarrow{P} \beta^*$.

With complete data from the finite population, we would solve the following estimating equation

$$\bar{U}(\beta) = \sum_{i \in \mathcal{P}} U_i(\beta) = \sum_{i \in \mathcal{P}} \frac{(y_i - \mu_i)}{V(\mu_i)} \left[\frac{\partial g(\mu_i)}{\partial \mu_i} \right]^{-1} x_i = 0.$$

which is an unbiased estimating equation for β_0 [46]. Then $\tilde{\beta}_N$ is the value of β that solves this equation. The design-based estimator [42, 16] that uses the sample solves the weighted estimating equation

$$\widehat{U}(\beta) = \sum_{i \in \mathcal{S}} w_i U_i(\beta) = 0.$$

to obtain $\widehat{\beta}_n$. This is an unbiased estimating equation for $\tilde{\beta}_N$ if $w_i \propto \pi_i^{-1}$. Assuming a law of large numbers and central limit theorem applies, then $\widehat{\beta}_n$ is asymptotically normal and consistent for β_0 if the super population model is correctly specified and for β^* in more general cases.

Variance estimators of coefficient estimates in survey regression models

The variance of the coefficient estimate obtained from a survey sample, $\widehat{\beta}_n$ is a sum:

$$\widehat{\text{Var}}(\widehat{\beta}_n) = \widehat{\text{Var}}_\pi(\widehat{\beta}_n) + \widehat{\text{Var}}_M(\tilde{\beta}_N) \tag{2.2}$$

where $\widehat{\text{Var}}_\pi$ is the sampling variance and $\widehat{\text{Var}}_M$ is the model based variance. The first term is of order n^{-1} and the second term is of order N^{-1} which can be ignored if the sample is much smaller than the finite population ($n << N$). The first term can estimated with a robust sandwich error derived using a delta-method argument [16] where

$$\widehat{\text{Var}}_\pi(\widehat{\beta}_n) = A^{-1}BA^{-1}.$$

A is the weighted observed Fisher's information matrix

$$A = \sum_{i \in \mathcal{S}} w_i \frac{\partial U_i(\beta)}{\partial \beta}\bigg|_{\beta=\widehat{\beta}_n}$$

and B is a consistent estimate of the variance of the design-based score function

$$B = \widehat{\text{Var}}_\pi\left[\widehat{U}(\beta)\right].$$

If the second term in Equation 2.2 is non-negligible it can be estimated with the estimated observed population Fisher's information matrix:

$$\widehat{\text{Var}}_M(\tilde{\beta}_N) = \sum_{i \in \mathcal{S}} \frac{\partial U_i}{\partial \beta}.$$

Variance estimators of coefficient estimates in survey regression models for convenience samples

As discussed earlier, sampling weights can be estimated for convenience samples. We can then use the estimated sampling weights to weight regression models to obtain inference about the target population. Using estimated sampling weights will likely lead to increased variability in the coefficients estimates and this needs to be accounted for in the variance estimate. In Chapter 3, we extend the variance estimate introduced above and account for uncertainty from estimating the sampling weights using a joint estimating equation approach similar to [103].

Methods for analyzing survey sample with inverse probability weights are very useful for obtaining estimates of population quantities. They can be extended for use in convenience samples using estimated sampling weights but there are three main issues that arise. First, you need an auxiliary representative data set to gain information about the target population. Second, you need to have a reasonable estimate of the sampling probabilities and thus the sampling weights. And third, you need to account for uncertainty from estimating the sampling weights in downstream parameter estimates. We address these three considerations in Chapter 3.

Inverse probability of sampling weights are useful in many contexts including estimates of causal effects and prediction error in the target population for samples with unequal sampling probabilities. We discuss methods for estimating causal effects and prediction error and how estimated sampling weights can be implemented and point out gaps in the existing methodology in the following sections.

2.2 Causal Inference

The goal of causal inference is to estimate or isolate causal effects. Colloquially, a causal effect is one where we can identify the effect of a given cause as opposed to an alternative explanation.[58]. Causation can be difficult to define, but Greenland, Robins and Pearl point to a definition from Hume in 1978. They say that event A caused event B if event A had to happen for event B to happen [50, 60]. In common terminology, we call event A the treatment or intervention. One example of a causal effect is the outcome of being exposed to a treatment versus the outcome of being exposed to the placebo[58].

2.2.1 Potential Outcomes

The framework discussed for causal inference and potential outcomes is based on Chapters 1 through 3 of Imbens and Rubin (2015) [62]. Following this textbook, we consider treatments with two groups: units who received an active treatment (treatment) and units who received a control treatment (control). These treatments are viewed symmetrically and can be thought of as different levels of the treatment. The framework discussed in this section can be extended to cases with more than 2 treatment groups or to a continuous treatment variable. Suppose we have a sample of n units and we have a treatment indicator A_i for unit i.

A unit level causal effect is a comparison in the outcome for a given unit at a given time if they were exposed to the treatment versus if they were exposed to the control. Consider unit i at a given time. The outcomes when being exposed to a specific treatment, for a particular unit, and at a particular point in time and space are called the potential outcomes. The potential outcome for unit i at a given point in time if they had received the treatment $(A_i = 1)$ is $Y_i(1)$ and the potential outcome if they had received the control $(A_i = 0)$ is $Y_i(0)$. An example of a unit level causal effect is

$$Y_i(1) - Y_i(0).$$

To compute a causal effect for a given unit we need to know their potential outcomes for both treatment options. "The fundamental problem of causal inference" is that we are only able to observe unit i in one treatment group at a given point in time and space [62, 58]. We are not able to observe all of the potential outcomes required to observe a unit level causal effect. The observed potential outcome for unit i under treatment A_i is

$$Y_i^{\text{obs}} = Y_i(A_i) = \begin{cases} Y_i(0) & \text{if } A_i = 0 \\ Y_i(1) & \text{if } A_i = 1 \end{cases}.$$

To estimate causal effects we need to impute the missing potential outcomes. To overcome "the fundamental problem of causal inference" and gain information about both treatment groups we need multiple observations. We can do this with multiple units at the same point in time or with one unit at multiple points in time.

2.2.2 Assumptions for Causal Inference

When using multiple units there may be alternative explanations besides a treatment effect for observed differences in the outcomes between treatment groups. We want to exclude these alternative explanations to isolate the causal effect. First, differences in the outcomes of each treatment group could be obscured by spillover effects where the treatment of one unit could impact the results of units from a different treatment group. Second, the differences between treatment groups could be caused by the reason that units end up in a given treatment group. To isolate the effect of the treatment, we need to prevent alternative explanations through assumptions or exclusion restrictions.

The Stable Unit Treatment Value Assumption (SUTVA)

The first assumption excludes spillover effects. The Stable Unit Treatment Value Assumption (SUTVA) is that the treatment assignment of a unit does not impact the potential outcomes of other units and there are no different forms of the treatment that lead to differences in potential outcomes [62]. The first part of SUTVA is often called "no interference" and attempts to exclude the existence of spillover effects. The second part is called the "no hidden variation of treatment" and excludes the possibility that differences in the treatment regiment within a treatment group are causing differences in potential outcomes.

Assumptions On The Assignment Mechanism

The second group of assumptions addresses how all units in the sample are selected for their treatment groups, which is the process called an assignment mechanism. Let W be an $n \times 1$ vector of assignments for the n units and X be covariates unrelated to the treatment. The assignment mechanism, $\Pr(W|X, Y(0), Y(1))$, is a function that assigns probabilities to each of the 2^n possible values of W. The assignment mechanisms must be in $[0, 1]$ and sum to one:

$$\sum_{W \in [0,1]^N} \Pr(W|X, Y(0), Y(1)) = 1$$

for all X, $Y(0)$, and $Y(1)$. Note that the assignment mechanism is not the probability of unit i receiving a specific treatment, but it is instead the probability of a full vector of assignments for all n units, W. Let W_i be the i-th value of W, then the unit-level assignment probability is

$$\Pr(A_i = 1|X, Y(0), Y(1)) = \sum_{W:W_i=1} \Pr(W|X, Y(0), Y(1)).$$

In randomized studies, we randomly assign each unit to a treatment group and thus control the treatment assignment. In observational studies, we do not control the treatment assignment. For example, in Chapter 3 we discuss an analysis of the effect of Vitamin E supplementation on measures of functional performance for older adults. The people in the study chose if they would take Vitamin E supplements or not. There are three assumptions for the assignment mechanism.

First, the individualistic assignment excludes the possibility that the treatment assignment of a given unit is influenced by the pre-treatment covariates or potential outcomes of other units. Definition 3.4 from Imbens and Rubin (2015)[62] states that given a function $q(\cdot) \in [0, 1]$, an assignment mechanism is individualistic if

$$\Pr(A_i = 1|X, Y(0), Y(1)) = q(X_i, Y_i(0), Y_i(1)), \qquad \forall \quad i \in 1, ...n \tag{2.3}$$

and

$$\Pr(W|X, Y(0), Y(1)) = c \prod_{i=1}^{n} q(X_i, Y_i(0), Y_i(1))^{W_i} (1 - q(X_i, Y_i(0), Y_i(1)))^{1-W_i} \tag{2.4}$$

for $(W, X, Y(0), Y(1)) \in \mathbf{A}$ for some set $\mathbf{A}$, and zero otherwise. The constant, c, is necessary so the assignment mechanisms sum to one. Individualistic assignment is generally considered a reasonable assumption.

Second, the probabilistic assignment assumptions requires that each unit has a non-zero probability of being in all treatment groups, or equivalently $\Pr(W|X, Y(0), Y(1))$ is probabilistic if

$$0 < \Pr(A_i = 1|X, Y(0), Y(1)) < 1 \tag{2.5}$$

for all $i \in 1, \ldots, n$.

Third, the unconfounded assumption requires that the potential outcomes are independent of assignment mechanism conditional on the observed covariates.

$$\Pr(W|X, Y(0), Y(1)) = \Pr(W|X). \tag{2.6}$$

Notice the conditioning on X in Equations 2.3, 2.4, 2.5, and 2.6. The assumptions are more reasonable when we condition on pre-treatment variables or covariates for each unit that are not influenced by the treatment. We often cannot verify these assumptions with data, however, so we need to use domain knowledge to assess if they are reasonable.

Given the individualistic assignment assumption, when the probabilistic assignment and unconfounded assumptions hold the treatment assignment is said to be strongly ignorable. If the treatment mechanism is a function of the treatment assignment, the covariates, and the observed potential outcomes then it is called ignorable.

2.2.3 Causal Estimands

There are different ways to define a causal effect. We could be interested in the average difference in outcome for the whole population. We could also be interested in the causal effect for a subpopulation such as the average difference in outcomes among males.

Average treatment effect (ATE)

The average treatment effect (ATE) is the average difference in outcomes comparing units in the treatment group to those in the control group for the population. It can be written as

$$E[Y(1) - Y(0)].$$

This is the definition of the ATE, but of course we do not observe both potential outcomes for each unit and can estimate it by imputing the missing potential outcomes.

Average treatment effect on the treated (ATT)

The average treatment effect on the treated (ATT) is the treatment effect among individuals who received treatment. It requires imputation of the potential outcome under control. It is written as

$$E[Y(1) - Y(0)|A = 1].$$

The average treatment effect on the controls (ATC) can be defined analagously.

Conditional average treatment effect (CATE)

The conditional average treatment effect (CATE) [1] is the average treatment effect at a set value of a covariate $X = x$:

$$E[Y(1) - Y(0)|X = x].$$

X_1 can be continuous or a factor. The CATE is useful for quantifying treatment effect heterogeneity if the treatment effect varies with X_1. Abrevaya et al. provide an example of the CATE where the effect of mother's smoking on the birth weight of babies among first time mothers is a function of the mother's age [1].

Local average treatment effect (LATE)

In a randomized clinical trial, non-compliers are participants who did not take their assigned treatment. We assume that compliance is all or nothing, participants either take all of the treatment or none of it. We also assume compliance is one-sided, it can only occur in the treatment group. The control group units cannot decided to not comply by starting to take the treatment if it is unavailable to them. The local average treatment effect (LATE) is the treatment effect among compliers (Chapter 23 of Imbens and Rubin [62]). Let G be an indicator of compliance, then

$$E[Y(1) - Y(0)|G = 1].$$

To estimate the LATE we need to use an instrumental variable. An instrumental variable is a variable that is almost certainly related to the treatment (A_i) but has no direct effect on the outcome outside of it's effect "through" the treatment. To estimate the LATE we will use the treatment assignment (as opposed to the receipt of treatment) as the instrumental variable.

There are two assumptions required for an instrumental variables approach. First, we need to assume that the treatment assignment is unconfounded, even though the receipt of treatment is confounded due to non-compliance. This assumption is satisfied in a randomized study. Second, we need an exclusion assumption that requires that the treatment assignment has no effect on the outcome for non-compliers. So, for people in the treatment group who did not take their assigned treatment we assume their original assignment does not impact their outcome. This is a reasonable assumption in randomized trials, but it is even more plausible if the study was double-blind and neither the investigators or the participants know the treatment assignment.

To estimate the LATE we need to estimate two intent-to-treat (ITT) effects. An intent-to-treat analysis compares the treatment effect among all people in a study based on their randomized treatment assignment, regardless of any noncompliance. The first effect is the ITT effect of the treatment assignment on the outcome. The second effect is the ITT effect of the treatment assignment on the receipt of treatment.

The LATE estimand is then equal to the ratio of the ITT effect of the treatment assignment on the outcome and the ITT effect of the treatment assignment on the receipt of treatment.

2.2.4 Estimating Causal Effects in Observational Studies

This subsection is based on Chapters 12 and 13 of Imbens and Rubin unless stated otherwise [62]. In observational studies, the probability of treatment assignment is an unknown function. Researchers conducting a study do not assign people to a treatment group. In observational studies, we still need the individualistic assignment mechanism, probabilistic assumption, unconfounded assumption, and SUTVA to estimate a causal effect.

Using covariates to make the unconfounded assumption reasonable is very important for observational studies. Let X_i be a vector of pretreatment covariates for unit i. The unconfounded assumption can be written as

$$Y_i(0), Y_i(1) \perp A_i \mid X_i.$$

Thus, it is important to carefully consider which variables might explain a spurious relationship between the treatment and potential outcomes and ensure that is collected and adjusted for. When the individualistic assignment mechanism, probabilistic assumption, unconfounded assumption, and SUTVA are satisfied, we can fairly compare the outcomes of units with the same values of X_i as we would in a randomized experiment. For example, if the unconfounded assumption held when conditioning on sex, then we could estimate the causal effect among females and males separately and do a weighted average of these estimates based on their subpopulation sizes to get an estimate of the ATE.

This estimation procedure could be expanded for higher dimension X_i, but at larger dimensions there will likely be subpopulations with zero units in one of the treatment groups. Instead, we could match units across the treatment and control groups with similar values of the covariates and estimate the treatment effect among matched pairs.

Propensity scores

To reduce the dimensionality of the problem and make matching easier, we can match on a scalar function of the covariates called a balancing score, $b(X_i)$ which is defined as a function of the covariates where

$$A_i \perp X_i \mid b(X_i).$$

X_i itself is a balancing score. The difference in outcomes for units exposed to treatment and control at a set value of the balancing score is unbiased for the treatment effect at that value. Thus, balancing on the propensity score can provide an unbiased estimate of the ATE [95].

One example of a scalar balancing score is the propensity score (Lemma 12.1 in Imbens and Rubin (2015) [62]). The propensity score is the probability of receiving treatment conditioned on the covariates or $e(X_i) = \Pr(A_i = 1|X_i = x_i)$. According to Lemma 12.2 in the same book [62], if treatment assignment is unconfounded, then it is unconfounded given a balancing score. Thus, if we condition on the propensity score we can estimate a causal effect. The distribution of the covariates across treatment groups should be the same given the propensity score.

The propensity score is the probability of a binary treatment indicator, so it can be estimated using binary prediction methods. Note that the propensity score definition can be expanded for continuous treatments [56]. For more discussion about prediction models, see Section 2.3. Alternatively, you can exploit the balancing properties of the propensity score and select the propensity score that best balances the covariate distributions across the treatment and control groups on propensity scores. For an example of how to estimate propensity scores see Chapter 13 of Imbens and Rubin (2015) [62]. In practice this is done by comparing the mean and variance of each covariate between the treatment and control groups within strata of the propensity scores. The approaches are commonly combined in an iterative approach where the propensity scores are estimated with a prediction model, then the covariate balance is checked, and then modifications to the prediction model are made as necessary and this process is repeated until the estimated propensity score gives good balance. This process involves updating a prediction model, but the outcome variable is not apart of it which prevents multiple testing which would inflate the Type I error (or false positive rate).

Once the propensity scores are estimated, they can be used to estimate the causal effect in several ways. First, propensity score stratification (or blocking) is an extension of matching units based on similar covariate values. We can assign units into statra based on their propensity score value, estimate the causal effect within each strata, and aggregate across strata.

Second, inverse probability weighting with the propensity score is when observations are weighted when estimating causal estimates. This approach is similar to IPW for sampling weights discussed in Section 2.1.3 but the weights are a function of the propensity score. There are different weighting schemes for different causal estimands. For example, the ATE can be estimated as follow:

$$\widehat{\text{ATE}} = \frac{1}{n}\sum_{i=1}^{n} \frac{A_i Y_i^{\text{obs}}}{e(X_i)} - \frac{1}{n}\sum_{i=1}^{n} \frac{(1-A_i)Y_i^{\text{obs}}}{1-e(X_i)}.$$

The ATT [69] can be estimated with:

$$\widehat{\text{ATT}} = \frac{1}{n}\sum_{i=1}^{n} A_i Y_i^{\text{obs}} - \frac{1}{n}\sum_{i=1}^{n} \frac{e(X_i)}{1-e(X_i)}(1-A_i)Y_i^{\text{obs}}.$$

A third option for accounting for the propensity score to estimate the causal effect is propensity score adjustment. We fit a regression model with the treatment indicator and the estimated propensity score as covariates and the outcome as the dependent variable. We can select a regression model based on the distribution of the outcome. A generalized linear model (GLM) [85] with link function $g(\cdot)$ is

$$g(E[Y_i^{\text{obs}}]) = \alpha_0 + \alpha_1 \cdot A_i + \alpha_2 \cdot \widehat{e}(X_i).$$

When $g(\cdot)$ is the identify link, the estimate of α_1 is an estimate of the ATE.

Covariate adjustment

An alternative approach for accounting for covariates in estimating a treatment effect is a regression analysis or covariate adjustment [3]. For this method, covariates are included as predictors in a regression model along with the treatment indicator and the outcome is still the dependent variable. A covariate adjusted model can take the form:

$$g(E[Y_i^{\text{obs}}]) = \beta_0 + \beta_1 A_i + \delta X_i^* \tag{2.7}$$

where X_i^* is a $1 \times q$ vector containing variables from X_i along with transformations of X_i and δ is a $q \times 1$ vector of parameters. The estimate of β_1 is an estimated causal effect. Covariate adjustment still requires the assumptions introduced in Section 2.2.2 (SUTVA, individualistic assignment, probabilistic assignment, and unconfoundedness) for making causal statements.

Comparing propensity scores and covariate adjustment

The main strength of covariate adjustment is that the model is very interpretable. It is easy to explain and show when a reasonable number of variables are adjusted for. For example, assuming an identity link function ($g(E[Y_i^{\text{obs}}] = E[Y_i^{\text{obs}}])$), β_1 in Equation 2.7 is interpreted as the mean difference in Y comparing people in the treatment group to those in the control group conditioning on all of the variables in X_i^* [105, 119]. It also requires the statistician to think about which covariates are being included as confounders which could prevent inclusion of variables that are independent of the response which increase the uncertainty in the treatment effect estimate [105]. A limitation of regression methods is that they rely on extrapolation if the covariate distributions in the treatment and control groups are dissimilar. Additionally, there can be confounding from misspecifying the functional form of the covariates in the model [119, 30, 31, 97].

An advantage of the propensity score is that the model estimating it can be very flexible and adjusted repeatedly without using information for the response. This reduces the temptation of adjusting the covariate adjusted regression model to get a significant result which inflates the Type I error rate [62]. The propensity score is also interpretable because it is easy to explain that the causal effect is estimated among units with a similar propensity for treatment [119]. A literature review of 47 observational studies using propensity scores published in 2001 assessed what details on the propensity score analysis were included in the paper. They found that 24 of the articles did not state how variable selection for the propensity score model was performed, 13 did not state the number of variables in the propensity score model, and 44 did not include information on the form of the covariates [114]. When researchers are not explicit about the methods and variables used to estimate the propensity score, it can be difficult to replicate estimates of causal effects in observational studies.

Both the propensity score and covariate adjustment can account for confounding. When estimating the propensity score analysts can select the covariates that minimize the prediction error and balance covariates, but covariate adjustment requires the analyst to think about which and how each covariates will be included in the model based on their relation to the treatment and response.

2.2.5 Causal Inference Under Non-Uniform Sampling

All of the methods discussed in this section assume the sample used to estimate the causal effect is representative of a target population. Estimating propensity scores and propensity-adjusted causal effects need to account for the sampling scheme. There is some debate in the literature about whether sampling weights should be included when estimating the propensity score [93, 119, 32, 9, 68]. To our knowledge, no one has quantified the impact of using sampling weights estimated for convenience samples when estimating propensity-adjusted causal effects. Additionally, there is need of a variance estimator of the estimated treatment effect that accounts for uncertainty in estimating the sampling weights, estimating the propensity score, and estimating the propensity-adjusted causal effect.

Propensity Scores for Convenience Samples

Previous papers have studied how to use propensity scores to estimate causal effects in survey samples with non-uniform, but prespecied sampling probabilities [93, 9, 68]. There are two steps for estimating a causal effect: (1) estimate the propensity score and (2) balance or match on the propensity score. There is some debate in the literature about whether sampling weights should be implemented in one or both of these steps. Ridgeway et al. (2015) [93] concludes that sampling weights should be included in both steps. In contrast Austin et al. (2018) [9] and Lenis et al. (2019) [68] suggest that only the outcome model (step 2) should be weighted, but they assume that the propensity score model is correctly specified.

There is room for further work on this problem and particularly for analyses of convenience samples where the sampling weights are not known. First, there is need of clarification about the necessity of weighting the propensity score model when it is not correctly misspecified, because it is unlikely to be correct in practice. Second, to our knowledge no one has assessed the impact of utilizing estimated sampling weights when using the propensity score to estimate causal effects. Furthermore, it is important to know how well the sampling weights need to be estimated to address sampling bias in estimated causal effects. Lastly, uncertainty from estimating both the sampling weights and the propensity score needs to be accounted for when quantifying the variability of causal effect estimates. We address these gaps in the research in Chapter 4.

2.3 Prediction Error

There are many options when selecting a predictive model. First, there are different classes of prediction models to choose between, including maximum likelihood methods such as generalized linear models, algorithmic methods such as random forest, shrinkage methods such as ridge regression, and ensemble methods that combine multiple models. Second, within one type of model you can also choose different degrees of complexity. For example, if you were using linear regression for prediction you will have to specify which covariates and transformations of the covariates will be included in the model. The standard way to define the "best" model is the model that generalizes the best to the target population that prediction model will be used for. Generalization is how well a given model performs on an independent test set from the target population. Most of the material through Section 2.3.7 (except Efron's general covariance penalty) is from Chapter 7 of the textbook "The Elements of Statistical Learning" [54]. References for where the methods were originally proposed are included when they are introduced.

2.3.1 Loss Functions

Generalization is quantified using a loss function. Let X be a vector of covariates, Y be the response, and $\widehat{f}(X)$ be the predicted value of Y using model $\widehat{f}(\cdot)$. A loss function, $L(Y, \widehat{f}(X))$, quantifies the difference between Y and $\widehat{f}(X)$. Some examples of commonly used loss functions when Y is continuous are absolute loss,

$$L_1(Y, \widehat{f}(X)) = |Y - \widehat{f}(X)|;$$

squared error loss,

$$L_2(Y, \widehat{f}(X)) = (Y - \widehat{f}(X))^2; \tag{2.8}$$

0-1 loss,

$$L_3(Y, \widehat{f}(X)) = I(Y \neq \widehat{f}(X));$$

and log-likelihood loss,

$$L_4(Y, \widehat{f}(X)) = -2 \times \text{loglik} \tag{2.9}$$

where loglik is the log-likelihood, or the joint probability distribution of Y [23]. The -2 term in the definition of the log-likelihood loss relates it to squared error loss when a Normal likelihood is assumed and $\sigma = \text{Var}(Y)$ is known. The Normal log-likelihood is

$$\log \Pr(Y|X) = \log\left(\frac{1}{\sigma\sqrt{2\pi}}\right) - \frac{1}{2}\left(\frac{Y - \widehat{f}(X)}{\sigma}\right)^2.$$

Notice that this is equivalent to Equation 2.9 since the first term is constant for all models.

2.3.2 Error Definitions

To quantify how well a model generalizes to an independent test set, we can consider the test error. The test error, $\mathrm{Err}_{\mathcal{T}}$, is also referred to as the generalization error or the extra-sample error because it is computed for a new sample (X^0, Y^0) drawn from the joint distribution of X and Y. Let $\mathcal{T}$ be the sample used to train the predictive model then

$$\mathrm{Err}_{\mathcal{T}} = E_{(X^0,Y^0)}\Big[L(Y^0, \widehat{f}(X^0))|\mathcal{T}\Big] \tag{2.10}$$

for fixed $\mathcal{T}$. This quantity is the test set for a given training set $\mathcal{T}$. The expected prediction error (EPE) or the expected test error averages over X, Y, and $\mathcal{T}$ and is equal to

$$\mathrm{EPE} = E_{\mathcal{T}}\Bigg\{E_{(X^0,Y^0)}\Big[L(Y^0, \widehat{f}(X^0))|\mathcal{T}\Big]\Bigg\}. \tag{2.11}$$

Ideally, we would like to pick the model that minimizes $\mathrm{Err}_{\mathcal{T}}$ in Equation 2.10 because it is the error in the target population given the training sample you used. This is difficult to estimate when you only have one sample, so we instead estimate EPE. An obvious estimate of the expected prediction error is the training error or the loss function applied directly to the training sample. Suppose there are n observations (x_i, y_i) in the training sample and $i = 1, \ldots, n$. The training error, err is

$$\mathrm{err} = \frac{1}{n}\sum_{i=1}^{n} L(y_i, \widehat{f}(x_i)). \tag{2.12}$$

The training error tends to underestimate EPE because the model is estimated and evaluated with the same sample. As models grow more complex they tend to adapt to idiosyncrasies in the sample. Some of the patterns in the sample may generalize to other samples from the target population but others may not. More complex models tend to have lower bias, but have higher variability across training samples. The training error will continue to decrease as the model complexity grows, but the EPE usually starts to increase at some point. As an extreme example, consider a model that is overfit to the point where $\widehat{f}(x_i) = y_i$. The training error will be 0 for the saturated model, but a new sample will most likely have new values of Y and so the EPE will be non-zero.

2.3.3 Model Selection and Assessment

There are two objectives for estimating prediction error. The first goal is model selection to compare between models of different complexities or types. This is operationalized by selecting the model with the smallest loss in an independent test set. When comparing models of different complexities, we can introduce a complexity parameter α and index the predictions from a model with complexity α as $\widehat{f_\alpha}(x_i)$. The second goal is model assessment for a given model, where you are interested in estimating the prediction error in an independent test set.

2.3.4 Strategies for Model Selection

Suppose you have P different models and you want to select the model $\widehat{f_{\alpha,p}}(x_i)$ where $p = 1, \ldots, P$ with the smallest EPE. One strategy for selecting a model is to estimate the EPE for all candidate models in your scope $\widehat{f_{\alpha,p}}(x_i)$ where $\{\alpha_1, ..., \alpha_P\}$ are the parameters for each of the P models. You then select the model with the smallest estimated EPE. This approach is termed best subsets. Best subsets can be computationally expensive because you have to fit P models and estimate the EPE P times. An alternative approach that reduces the computational burden is to pick a subset of the candidate models to fit the model in and compute the prediction error for and choose the best model among the subset. One framework for doing this is stepwise selection in regression. For forward stepwise selection you start with the model that only contains an intercept. Then you compare all models with one covariate added and select the model with the smallest estimated prediction error. You repeat this process until there are no covariates left that decrease the estimated prediction error. Another option is backwards selection where you start with the largest model and remove one covariate at a time. For a comparison between best subsets and stepwise forward selection see Hastie (2020) [55].

Bias variance trade off

When selecting a model, you have to make a trade off between bias and variance. Complex models will fit the training data better and have lower bias, but high variance in the test set. Simpler models will not fit the data as closely but will have lower variance. Consider the case of additive errors, where $Y = f(X) + \epsilon$, $E(\epsilon) = 0$, and $\mathrm{Var}(\epsilon) = \sigma_\epsilon^2$. Under squared error loss we can write down the relationship between EPE, bias, and variance for a regression fit $\widehat{f}(X)$ evaluated at $X = x_0$.

$$\begin{aligned}
\mathrm{EPE}(x_0) &= E\Big[\big(Y - \widehat{f}(x_0)\big)^2 | X = x_0\Big] \\
&= E\Big[\big((Y - E[\widehat{f}(x_0)]) + (E[\widehat{f}(x_0)] - \widehat{f}(x_0))\big)^2 | X = x_0\Big]
\end{aligned}$$

We start by expanding this expression and let K be the product of the first and second terms. We will show that $K = 0$ later. So,

$$\begin{aligned}
&= E\Big[(Y - E[\widehat{f}(x_0)])^2 + (E[\widehat{f}(x_0)] - \widehat{f}(x_0))^2 | X = x_0\Big] + 2K \\
&= E\Big[(Y - E[\widehat{f}(x_0)])^2 + (E[\widehat{f}(x_0)] - \widehat{f}(x_0))^2 | X = x_0\Big] + 0 \\
&= E\Big[(Y^2 - 2YE[\widehat{f}(x_0)] + E[\widehat{f}(x_0)]^2) | X = x_0\Big] + V
\end{aligned}$$

Where $V = E\Big[(E[\widehat{f}(x_0)] - \widehat{f}(x_0))^2 | X = x_0\Big]$. Now, we can replace Y with $f(X) + \epsilon$ and evaluate the expectation,

$$\begin{aligned}
&= E\Big[f(x_0)^2 + \epsilon^2 + 2f(x_0)\epsilon - 2f(x_0)E[\widehat{f}(x_0)] - 2\epsilon E[\widehat{f}(x_0)] + E[\widehat{f}(x_0)]^2 | X = x_0\Big] + V \\
&= f(x_0)^2 + \sigma^2_\epsilon + 0 - 2f(x_0)E[\widehat{f}(x_0)] - 0 + E[\widehat{f}(x_0)]^2 + V \\
&= \sigma^2_\epsilon + \big(f(x_0)^2 - 2f(x_0)E[\widehat{f}(x_0)] + E[\widehat{f}(x_0)]^2\big) + V \\
&= \sigma^2_\epsilon + \big(f(x_0) - E[\widehat{f}(x_0)]\big)^2 + E\Big[(\widehat{f}(x_0) - E[\widehat{f}(x_0)])^2 | X = x_0\Big]
\end{aligned}$$

Multiplying the third term by $(-1)^2$ gives us,

$$\text{EPE}(x_0) = \sigma^2_\epsilon + \text{Bias}^2(\widehat{f}(x_0)) + \text{Var}(\widehat{f}(x_0)) \tag{2.13}$$

Finally, we will show that the cross-term K is equal to 0. Note that $Y \perp \widehat{f}(x_0)$ because the predicted value is a function of the training set and the new observation Y is from an independent sample.

$$\begin{aligned}
K &= E\Big[(Y - E[\widehat{f}(x_0)])(E[\widehat{f}(x_0)] - \widehat{f}(x_0)) | X = x_0\Big] \\
&= E\Big[(Y - E[\widehat{f}(x_0)]) | X = x_0\Big] \times E\Big[(E[\widehat{f}(x_0)] - \widehat{f}(x_0)) | X = x_0\Big] \\
&= 0
\end{aligned}$$

Thus, the EPE is the sum of the irreducible variance σ^2_ϵ, bias squared and the variance in this scenario. Reducing bias or variance will necessarily inflate the other. Additionally, we can write out the variance term if we assume that the model is linear and fit with least squares, or $\widehat{f_p}(x) = x^T\widehat{\beta}$ where β is a $p \times 1$ vector of parameter estimates and p is the number of covariates. Let $\mathbf{X}$ be the $n \times p$ matrix of predictors where row i corresponds to observation

i and $\mathbf{y}$ is the $p \times 1$ vector of observations, then

$$\begin{aligned}\text{Var}(\widehat{f}(x_0)) &= \text{Var}(x_0^T \widehat{\beta}) = \text{Var}(x_0^T (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}) \\ &= x_0^T (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T \text{Var}(\mathbf{y})\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}x_0 = \sigma_\epsilon^2 x_0^T (\mathbf{X}^T\mathbf{X})^{-1}x_0.\end{aligned}$$

Notice that this expression is a function of x_0. If we set x_0 equal to the observed values of x_i we can take an average across x_i. Using the definition of the trace and the commutativity property of the trace,

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n} \sigma_\epsilon^2 x_i^T (\mathbf{X}^T\mathbf{X})^{-1}x_i &= \frac{1}{n}\sigma_\epsilon^2 \text{trace}\Big(\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\Big) = \frac{1}{n}\sigma_\epsilon^2 \text{trace}\Big((\mathbf{X}^T\mathbf{X})^{-1}(\mathbf{X}\mathbf{X}^T)\Big) \\ &= \frac{1}{n}\sigma_\epsilon^2 \text{trace}\Big(I_{p\times p}\Big) = \sigma_\epsilon^2 \frac{p}{n}.\end{aligned}$$

We can use this relationship with the average of Equation 2.13 to show

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n} EPE(x_i) &= \sum_{i=1}^{n}\Big(\sigma_\epsilon^2 + \text{Bias}^2(\widehat{f}(x_0)) + \text{Var}(\widehat{f}(x_0))\Big) \\ &= \sigma_\epsilon^2 + \frac{1}{n}\sum_{i=1}^{n}\Big(f(x_i) - E[\widehat{f}(x_i)]\Big)^2 + \sigma_\epsilon^2 \frac{p}{n}.\end{aligned}$$

Thus when fitting a linear model when the data follows the additive error model, the average EPE is a function of the number of parameters in the model.

In these examples, we can see that there is a trade-off between bias and variance when selecting a model and that prediction error tends to increase with model complexity.

Strategies for estimating prediction error

One strategy for estimating prediction error when you have a large enough data set is to split the sample into training, validation and test sets. The training sample is used for training the models, the validation sample is used for model selection, and the test set is used for model assessment. Hastie, Tibshirani, and Freedman [54] suggest using 50% of the data for training and 25% for both the validation and test samples. If using 50% of the data set for the training set will lead to increased bias because it is too small, you can split it into two samples: a training set and a combined validation/test set. This will lead to an underestimation of the test error because you are picking the model with the lowest test error which may be the lowest due to random chance.

In the following sections, I will discuss methods for estimating the prediction error for data sets that are too small to be split into separate training and validation sets. There are two categories of estimates of the prediction error: (1) analytic methods that try to directly estimate the difference between the training error and the test error and (2) resampling methods that resample the data to mimic having a validation set.

2.3.5 Analytic Methods

Analytic methods estimate prediction error by estimating the difference between the prediction error for a new sample and the training error. It is difficult to express this analytically for the EPE, so instead we can target the in-sample error ($\mathrm{Err}_{\mathrm{in}}$). The in-sample error holds the covariate distribution of x constant, but samples n new values of the response Y^0 at each x_i, so

$$\mathrm{Err}_{\mathrm{in}} = E_{\mathcal{T}}\Big\{E_{Y^0}\Big[L(Y^0, \widehat{f}(x_i))|\mathcal{T}\Big]\Big\} \tag{2.14}$$

The in-sample error is useful for model selection, but it is not usually of interest for model assessment because in practice the X values are likely to change for a new sample. Thus, analytic estimates of prediction error are more useful for model selection.

Optimism

Optimism (op) is the difference between the in-sample error and the training error,

$$\text{op} \equiv \text{Err}_{\text{in}} - \text{err} \tag{2.15}$$

Optimism tends to be positive since the training error usually under estimates the prediction error. It is easier to estimate the average optimism, ω, where the expectation is taken over y where,

$$\omega \equiv E_y(\text{op}).$$

For many loss functions, the expected optimism is a function of the covariance between y_i and the fitted value $\widehat{f}(x_i)$, or how strongly y_i determines it's own prediction.

$$\omega = \frac{2}{n}\sum_{i=1}^{n} \text{Cov}\big(y_i, \widehat{f}(x_i)\big) \tag{2.16}$$

The proof for squared error loss is as follows,

$$\begin{aligned}\omega &= E_y\Big\{\frac{1}{n}E_{Y^0}\Big[(Y_i^0 - \widehat{f}(x_i))^2\Big] - \frac{1}{n}(y_i - \widehat{f}(x_i))^2\Big\}\\ &= \frac{1}{n}\sum_{i=1}^{n}\Big(E_y[E_{Y^0}[(Y_i^0)^2]]\text{- 2E}_y[E_{Y^0}[Y_i^0\widehat{f}(x_i)]] + E_y[E_{Y^0}[(\widehat{f}(x_i))^2]]\\ &\qquad\qquad\text{- E}_y[y_i^2] + 2E_y[y_i\widehat{f}(x_i)] - E_y[(\widehat{f}(x_i))^2]\Big)\end{aligned}$$

Since $\widehat{f}(x_i) \perp Y_i^0$, $E_y[E_{Y^0}[(\widehat{f}(x_i))^2]] = E_y[(\widehat{f}(x_i))^2]$ and the third and sixth terms cancel each other out. Additionally, $y_i \sim D$ and $Y_i^0 \sim D$ for some distribution D but are independent. Thus, $E_y[E_{Y^0}[(Y_i^0)^2]] = E_y[(Y_i^0)^2] = E_y[y_i^2]$ and the first and fourth terms cancel. We are left with ,

$$\omega = \frac{2}{n}\sum_{i=1}^{n}\Big(E_y[y_i\widehat{f}x_i] - E_y[E_{Y^0}[Y_i^0\widehat{f}(x_i)]]\Big)$$

We can use the properties $\widehat{f}(x_i) \perp Y_i^0$ and $E_y[Y_i^0] = E_y[y_i]$ again to obtain the final result,

$$\begin{aligned}\omega &= \frac{2}{n}\sum_{i=1}^{n}\Big(E_y[y_i\widehat{f}(x_i)] - E_y[y_i]E_y[\widehat{f}(x_i)]\Big)\\ &= \frac{2}{n}\sum_{i=1}^{n}\text{Cov}\big(y_i, \widehat{f}(x_i)\big)\end{aligned}$$

When $\widehat{f}(x_i)$ is a linear function of y or $\widehat{f}(x_i) = \sum_{j=1}^{n} c_j y_j$, Equation 2.16 simplifies further. Recall that $y_i \perp y_j$ when $i \neq j$, then

$$\begin{aligned}\omega &= \frac{2}{n}\sum_{i=1}^{n}\text{Cov}\big(y_i, \widehat{f}(x_i)\big) = \frac{2}{n}\sum_{i=1}^{n}\text{Cov}\Bigg(y_i, \sum_{j=1}^{n} c_j y_j\Bigg)\\ &= \frac{2}{n}\sum_{i=1}^{n}\text{Cov}\big(y_i, c_i y_i\big) = \frac{2}{n}\sum_{i=1}^{n} c_i\text{Var}(y_i).\end{aligned}$$

In least squares, c_i is the i-th diagonal element of the hat matrix. Under the additive variance model where $y_i = f(x_i) + \epsilon$ and $\text{Var}(\epsilon) = \sigma_\epsilon^2$,

$$\omega = \frac{2}{n}p\sigma_\epsilon^2 \tag{2.17}$$

where p is the number of parameters. Thus, the optimism increases with model complexity but decreases as the sample size increases.

A strategy for estimating the optimism and thus the in-sample error is to find an estimate of the covariance between y_i and $\widehat{f}(x_i)$. So analytic estimates of the in-sample error often take the form,

$$\widehat{\text{Err}}_{\text{in}} = \text{err} + \widehat{\omega}$$

Mallows C_p

Equation 2.17 leads to an estimate of the in-sample error for linear models with p parameters fit with squared error loss called Mallows C_p [78]:

$$C_p = \text{err} + \frac{2}{n} p \widehat{\sigma}_\epsilon^2$$

where $\widehat{\sigma}_\epsilon^2$ is estimated with the least biased (or most complex) model in the scope of consideration. As is common with many analytic estimates of Err_{in}, this takes the form of the training error plus a penalty term which in this case is a function of the number of parameters.

Akaike's information criterion (AIC)

Akaike's information criterion (AIC) [4, 5] is an extension to C_p for log-likelihood loss. Let $\widehat{y_i} = \widehat{f}(x_i)$ and $\widehat{\theta}$ be the maximum likelihood estimates of θ, a vector of parameters. Then the Err_{in} under log-likelihood loss is

$$\text{Err}_{\text{in}} = E_Y[-2 \log \text{Pr}_{\widehat{\theta}}(Y|X)]. \tag{2.18}$$

AIC depends on an asymptotic relationship that applies as $n \rightarrow \infty$ similar to Equation 2.17 used for Mallows C_p.

$$-2E_Y[\log \text{Pr}_{\widehat{\theta}}(Y|X)] \approx -\frac{2}{n} E_y[\text{loglik}] + 2\frac{p}{n}$$

where loglik $= \sum_{i=1}^{n} \log \Pr_{\widehat{\theta}}(y_i|X)$. Multiplying by n, the AIC estimate is

$$\mathrm{AIC} = -2\mathrm{loglik} + 2p. \tag{2.19}$$

This definition is equivalent to Mallows C_p for a Normal log-likelihood so C_p with known variance σ^2.

For a more general definition of AIC, p can be replaced by other measures of model complexity, such as the effective degrees of freedom. This is especially useful for regularization. Consider a linear model of the form,

$$\widehat{\mathbf{y}} = \mathbf{S}\mathbf{y}$$

where $\mathbf{S}$ is an $n \times n$ matrix that is a function of $\mathbf{X}$ but not $\mathbf{y}$. The effective degrees of freedom for a linear model are

$$\mathrm{df}(\mathbf{S}) = \mathrm{trace}(\mathbf{S}).$$

If $\mathbf{S}$ is an orthogonal-projection matrix onto a basis set spanned by p features, such as the hat matrix in ordinary least squares, then $\mathrm{trace}(\mathbf{S}) = p$. Replacing p with the effective degrees of freedom allows AIC to be used with a broader range of models.

AIC can also be derived by picking the candidate model with the smallest Kullback-Leibler divergence from the true model [5, 76, 28].

Efron's general covariance penalty

Efron proposed a class of analytic approximations to the in-sample error that allow for the use of general loss functions beyond log-likelihood loss [36]. He extends the relationship between optimism and covariance to the settings where y is generated from an unknown probability distribution, f. Efron's covariance penalty estimate applies for loss functions from the *q class* of error measures where $q(\cdot)$ is a concave function. Let, $\dot{q}(\widehat{y}) = \frac{\partial q}{\partial y}|_{y=\widehat{y}}$. Then the relationship between the loss function and $q(\cdot)$ is

$$L(y, \widehat{y}) = q(\widehat{y}) + \dot{q}(\widehat{y})(y - \widehat{y}) - q(y).$$

Let $\widehat{\lambda}_i = -\dot{q}(\widehat{y}_i)/2$. Efron showed that

$$\omega_i = 2\mathrm{Cov}(\widehat{\lambda}_i, y_i).$$

We will start with the definition of the err and $\mathrm{Err}_{\mathrm{in}}$ for a loss function from the *q class*:

$$\mathrm{Err}_{\mathrm{in},i} = E_{Y^0}\left[L(Y_i^0, \widehat{y}_i)\right] = q(\widehat{y}_i) + \dot{q}(\widehat{y})(E_{Y^0}[Y_i^0] - \widehat{y}_i) - E_{Y^0}[q(Y_i^0)]$$

$$\mathrm{err}_i = L(y_i, \widehat{y}_i) = q(\widehat{y}_i) + \dot{q}(\widehat{y})(y_i - \widehat{y}_i) - q(y_i).$$

Combining these expressions with the definition of optimism (Equation 2.15), we get

$$\begin{aligned}\mathrm{op}_i &= \mathrm{Err}_{\mathrm{in},i} - \mathrm{err}_i \\ &= \dot{q}(\widehat{y})\Big[(E_{Y^0}(Y_i^0) - \widehat{y}_i) - (y_i - \widehat{y}_i)\Big] - E_{Y^0}[q(Y_i^0)] + q(y_i) \\ &= 2\widehat{\lambda}_i(y_i - E_{Y^0}(Y_i^0)) - E_{Y^0}[q(Y_i^0)] + q(y_i).\end{aligned}$$

Since y_i and Y_i^0 follow the same distribution, $E[q(y_i)] = E[q(Y_i^0)]$ and $E[y_i] = E[Y_i^0]$. So the expected optimism,

$$\begin{aligned}
\omega_i &= E_y[\text{op}_i] \\
&= E_y\left[2\widehat{\lambda}_i(y_i - E_{Y^0}(Y_i^0))\right] - E_{Y^0}[q(Y_i^0)] + E_y[q(y_i)] \\
&= 2E_y\left[\widehat{\lambda}_i(y_i - E_y[y_i]\right] \\
&= 2(E_y[\widehat{\lambda}_i y_i] - E_y[\widehat{\lambda}_i]E_y[y_i]) \\
&= 2\text{Cov}(\widehat{\lambda}_i, y_i).
\end{aligned}$$

We can multiply by n to match the definition of AIC and obtain Efron's estimate of Err_{in}

$$\widehat{\text{Err}}_{\text{in}} = \sum_{i=1}^{n}(err_i + 2cov(\widehat{\lambda}_i, y_i)).$$

The covariance term can be estimated with a parametric bootstrap when there is not an analytic form available for a given loss function.

Bayesian information criterion (BIC)

The Bayesian information criterion (BIC) [104] is another estimate of the in-sample error for models fit by maximizing the log-likelihood, but it is motivated by a Bayesian framework for model selection.

Suppose we have M candidate models: $\{\mathcal{M}_1, \ldots, \mathcal{M}_M\}$ each with prior probability of being the true model $\Pr(\mathcal{M}_m)$, where $m = 1, \ldots, M$ and $\sum_{m=1}^{M} \Pr(\mathcal{M}_m) = 1$. Additionally, we have training data (x_i, y_i) where $i = 1, \ldots, n$. Let θ_m be the parameters for $\mathcal{M}_m$.

We want to select the model with the largest posterior probability. The posterior probability for model $\mathcal{M}_m$ is

$$\begin{aligned}\Pr(\mathcal{M}_m|\text{Data}) &\propto \Pr(\mathcal{M}_m)\Pr(\text{Data}|\mathcal{M}_m)\\ &\propto \Pr(\mathcal{M}_m)\int \Pr(\text{Data}|\mathcal{M}_m,\theta_m)\Pr(\theta_m|\mathcal{M}_m)d\theta_m\end{aligned}$$

It is common to assume a uniform prior probability on all models so $\Pr(\mathcal{M}_m) = \frac{1}{M}$ for all m. We can use a Laplace approximation to the integral in the above expression with some simplifications to obtain,

$$\Pr(\text{Data}|\mathcal{M}_m) = \log \Pr(\text{Data}|\widehat{\theta}_m, \mathcal{M}_m) - \frac{p_m}{2}\log n + O(1)$$

where $\widehat{\theta}_m$ is the maximum likelihood estimate of θ_m. Under the following loss function,

$$-2\log \Pr(\text{Data}|\widehat{\theta}_m, \mathcal{M}_m)$$

the BIC is

$$\text{BIC} = -2\text{loglik} + p\log(n).$$

Selecting the model with the smallest BIC is equivalent to selecting the model with the largest posterior probability. If you compute the BIC for each model, you can use it to compute the posterior probability. This is useful because comparing the posterior probabilities allows you to quantify the relative performance of each model. The posterior probability of model $\mathcal{M}_m$ is

$$\frac{e^{-\frac{1}{2}\text{BIC}_m}}{\sum_{\ell=1}^{M} e^{-\frac{1}{2}\text{BIC}_\ell}}.$$

AIC and BIC have similar penalities, but BIC replaces 2 with $\log n$. This means that BIC places a heavier penalty on complex models when $n > e^2 \approx 7.4$ and will tend to chose simpler models than AIC. BIC, however, is asymptotically consistent so if the model scope includes the true model, it will select the true model with probability approaching one as $n \to \infty$.

2.3.6 Resampling Methods

An alternative approach to analytic estimates of the optimism for estimating prediction error are resampling methods. Resampling methods resample the data to create separate training and evaluation sets. Resampling methods are more computationally intensive than analytic methods, but they directly estimate the EPE since they resample both x and y.

Cross-Validation (CV)

Cross-validation [106, 107, 6] mimics splitting the sample into training and test sets. K-fold cross-validation (CV) splits the data into K subsamples termed folds of approximately equal size. For each of the $k = 1, ..., K$ folds, the model is trained on the $K - 1$ other folds and validated on the k-th fold. Figure 2.1 illustrates the case when $K = 5$. A row denotes the $k - th$ iteration where the model is fit on the 4 other folds and the loss function is computed for the k-th fold. The loss function is combined across all folds.

More formally, let $k(i)$ be the fold that observation i is apart of and $\widehat{f}^{-k(i)}(x_i)$ denote the predicted value for y_i from the model fit on all folds besides $k(i)$. Then the K-fold CV statistic is

$$\mathrm{CV}_K = \frac{1}{n}\sum_{i=1}^{n} L(y_i, \widehat{f}^{-k(i)}(x_i)) \tag{2.20}$$

There are several considerations for how to pick K. When K approaches n there is substantial

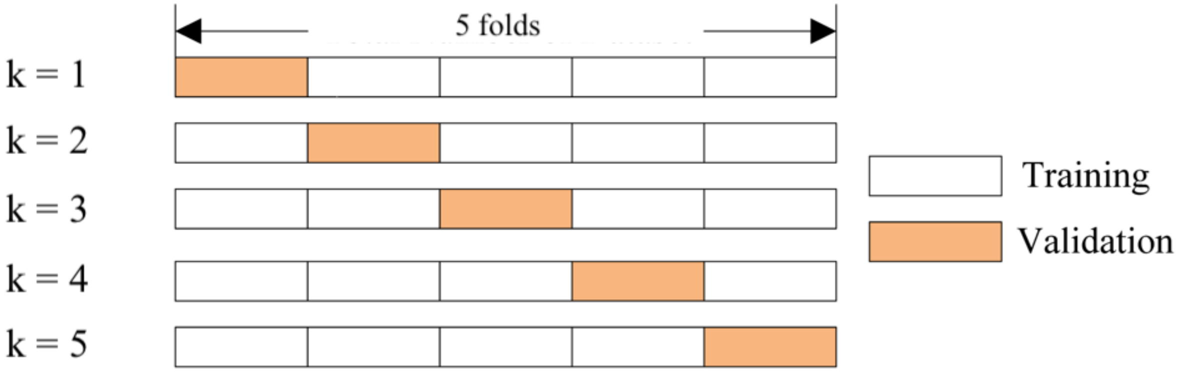


Figure 2.1: Example of 5-fold cross-validation. Each row represents the k-th iteration of the algorithm where the model ($k = 1, \dots, K$) is fit on all folds except k and the loss is computed for the k-th fold. Image is adapted from `https://www.kaggle.com/dansbecker/cross-validation`.

overlap between the training sets which leads to higher variability of the EPE estimate. For smaller K, the training sets are smaller because the fold size is roughly $n - n/K$ which leads to more bias. $K = 5$ or $K = 10$ is suggested as a compromise. When $K = n$, the so called Leave-One-Out CV (LOO CV), the CV statistic is approximately unbiased for the true EPE, but is highly variable. It is computationally expensive because it requires the model to be fit n times.

There is an analytic form for the LOO CV statistic for simple linear regression. It is based on the leave-one-out residuals that are the residuals that would be obtained if the model was fit without y_i [6]. Let $\widehat{y}_{(i)}$ be the predicted value of y_i from the model fit with all observations besides y_i, then the leave-one-out residual for y_i is,

$$y_i - \widehat{y}_{(i)} = \frac{y_i - \widehat{y}_i}{1 - \mathbf{H}_{ii}}$$

where $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T)$ is the hat matrix. The LOO CV estimator under squared error loss is,

$$\mathrm{CV}_{\mathrm{LOO}} = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{y_i - \widehat{y}_i}{1 - \mathbf{H}_{ii}}\right]^2.$$

This definition can be generalized by replacing the diagonal elements $\mathbf{H}_{ii}$ with their average, the effective degrees of freedom. The generalized CV estimate [47, 111] is

$$\mathrm{CV}_{\mathrm{GCV}} = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{y_i - \widehat{y}_i}{1 - \mathrm{trace}(\mathbf{H})}\right]^2.$$

Bootstrap (BS)

Bootstrapping is a general resampling technique useful for obtaining a distribution for sample statistics [34, 37]. Suppose we have data $\mathbf{Z} = \{z_1, ..., z_n\}$ where $z_i = (x_i, y_i)$. Bootstrapping involves sampling from Z with replacement B times to obtain B bootstrap samples of size n. You can fit the model of interest in each of the $b = 1, ..., B$ bootstrap samples, obtain a sample statistic such as a coefficient estimate and compute a Monte Carlo estimate of it's variance across bootstrap samples. The bootstrapped distribution should converge to the empirical distribution which converges to the true distribution.

A naive bootstrap estimate of the EPE would be to fit the model on each of the B bootstrap sample and compute the loss function for the original full sample. Let $\widehat{f}^{*b}(x_i)$ denote the predicted value for y_i from the model fit on the b-th bootstrap sample. The naive bootstrap estimate is

$$\mathrm{BS}_{\mathrm{naive}} = \frac{1}{B}\frac{1}{n}\sum_{b=1}^{B}\sum_{i=1}^{n} L(y_i, \widehat{f}^{*b}(x_i)).$$

There is substantial overlap between each bootstrap and the full sample so this estimate will likely underestimate the EPE. To avoid this, for each bootstrap sample b, we can fit the model on the bootstrapped sample and calculate the loss function for all z_i not included in the b-th bootstrap sample. Let C_{-i} be the set of indices of bootstrap samples that do not include z_i and $|C_{-i}|$ be the size of the set. Then the leave-one-out bootstrap estimate of the

EPE is,

$$\mathrm{BS}_{\mathrm{LOO}} = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{|C_{-i}|}\sum_{b\in C_{-i}} L(y_i, \widehat{f}^{*b}(x_i)).$$

This estimate runs into the small sample size problem discussed for K-fold CV when K is small. The probability an observation is contained in bootstrap sample b is,

$$\begin{aligned}\Pr(z_i \in \text{ BS sample } b) &= 1 - \Big(1 - \frac{1}{n}\Big)^n \\ &\approx 1 - e^{-1} \\ &= 0.632.\end{aligned}$$

So the number of unique observations in bootstrap sample b is approximately $0.632n$. Thus, the LOO BS estimator will likely overestimate the EPE. To address this, the .632 bootstrap estimator pulls the LOO estimate down towards the training error using a weighted average. The .632 estimator [35] is

$$\mathrm{BS}_{.632} = .368 \cdot \mathrm{err} + .632 \cdot \mathrm{BS}_{\mathrm{LOO}}. \tag{2.21}$$

This estimator breaks down in overfit situations. Take for example, the saturated model with err $= 0$, the .632 estimate is

$$\mathrm{BS}_{.632}(\text{saturated model}) = .368 \cdot 0 + .632 \cdot \mathrm{BS}_{\mathrm{LOO}} = .632 \cdot \mathrm{BS}_{\mathrm{LOO}}$$

which will underestimate the EPE. One solution is to adjust the weighting scheme in Equation 2.21 to account for the degree to which the model is overfit. The no-information error rate is the error rate when the predictors are independent of the response which can be estimated with the permutation distribution by computing the loss function for all possible combinations of $\mathbf{X}$ and $\mathbf{y}$. The estimate of the no-information error rate, $\widehat{\gamma}$ is

$$\widehat{\gamma} = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} L(y_i, \widehat{f}(x_j)).$$

We can use this to estimate the relative overfitting rate:

$$\widehat{R} = \frac{\mathrm{BS}_{\mathrm{LOO}} - \mathrm{err}}{\widehat{\gamma} - \mathrm{err}}.$$

If there is no overfitting and $\mathrm{BS}_{\mathrm{LOO}} = \mathrm{err}$, $\widehat{R} = 1$ but if the LOO BS estimate is equal to the no-information error rate ($\mathrm{BS}_{\mathrm{LOO}} = \widehat{\gamma}$) $\widehat{R}$ will be 0. Then the .632+ bootstrap estimator [38] is

$$\mathrm{BS}_{.632+} = (1 - \widehat{w}) \cdot \mathrm{err} + \widehat{w} \cdot \mathrm{BS}_{\mathrm{LOO}}$$

where

$$\widehat{w} = \frac{.632}{1 - .368\widehat{R}}.$$

The .632+ BS estimate ranges from the .632 BS estimate when $\widehat{R} = 0$ to the LOO BS estimate when $\widehat{R} = 1$.

2.3.7 Comparing Analytic and Resampling Methods

Resampling methods are more computationally intensive than analytic methods, but they estimate the expected prediction error (EPE) instead of the in-sample error ($\mathrm{Err}_{\mathrm{in}}$). If your goal is model assessment, resampling methods will give a better estimate of the EPE which is usually of interest since the y values will likely change for a new sample. Analytic methods are more useful for model selection when estimating the prediction error is only necessary for comparing between models. Resampling methods are more difficult for model selection because it has to be computed for each candidate model. Another advantage of resampling methods is they can be computed for any loss function, but Efron's analytic covariance penalty estimate works for a broad class of loss functions. Stone (1977) [107] showed that AIC and leave-one-out cross-validation are asymptotically equivalent assuming the true model is in the scope.

2.3.8 Estimates of Prediction Error Under Non-Uniform Sampling

The prediction assessment methods introduced above assume that training sample is representative of the target population. When observations are sampled with unequal sampling probabilities, as is the case for survey samples or convenience samples these methods will not be a good estimate of the prediction error in an independent simple random sample from the target population. They will instead estimate the prediction error for an independent sample drawn with the same sampling scheme as the training sample. There are analogues to analytic and resampling prediction error estimates that account for the sampling scheme by including sampling weights.

Design-based AIC

Lumley and Scott [76] extended AIC and BIC to include sampling weights. The derivation of the design-based AIC, or dAIC follows the derivation under uniform sampling from [28]. The derivation is motivated by estimating and minimizing the Kullback-Leibler (KL) divergence of the candidate model from the true model. Let w_i be the sampling weight for observation i under the constraint that $\sum_{i=1}^{n} w_i = 1$. Suppose that the true model is $f(x)$ and an estimated candidate model is $\widehat{f}_\theta(x)$ with parameters θ. The KL divergence between the true model and the estimated model is

$$KL(f, \widehat{f}_\theta) = E_f\left[\log\left\{\frac{f(x)}{\widehat{f}_\theta(x)}\right\}\right] = E_g\big[\log f(x)\big] - \ell(\theta)$$

where $\ell(\theta) = E_f[\log \widehat{f}_\theta(x)]$ is the expected log-likelihood in the target population. Only the second term is a function of θ so selecting θ that maximizes the log-likelihood is equivalent to minimizing the KL divergence. The maximum is obtained at a unique point θ^*. A sample estimate of $\ell(\theta)$ is the weighted sample mean,

$$\widehat{\ell}(\theta) = \sum_{i=1}^{n} w_i \ell\big(\widehat{f}_\theta(x_i)\big) = \sum_{i=1}^{n} w_i \ell_i(\theta).$$

Then $\widehat{\theta}$ is the value of θ that maximizes $\widehat{\ell}(\theta)$. Under some regularity conditions (see Fuller (2009) [42]) $\widehat{\theta}$ is consistent for θ^* as $n, N \rightarrow \infty$ where N is the population size. We also assume the asymptotic framework that there is a sequence of finite populations that are random samples from the target superpopulation. Let the population score be

$$U(\theta) = \frac{\partial \ell(\theta)}{\partial \theta}$$

and the weighted estimate of the score is

$$\widehat{U}(\theta) = \frac{\partial \widehat{\ell}(\theta)}{\partial \theta}$$

with the corresponding score equations $U(\theta^*) = 0$ and $\widehat{U}(\widehat{\theta}) = 0$. Under the regularity conditions,

$$\sqrt{n}(\widehat{\theta} - \theta^*) \xrightarrow{d} N(0, V(\theta^*)) \text{ as } n \rightarrow \infty.$$

The asymptotic covariance of $\sqrt{n}\widehat{\theta}$ can be estimated with a sandwich estimate:

$$\widehat{V}(\widehat{\theta}) = \widehat{\mathcal{I}}(\widehat{\theta})^{-1}\widehat{V}_U(\widehat{\theta})\widehat{\mathcal{I}}(\widehat{\theta})^{-1} \tag{2.22}$$

where $\widehat{V}_U(\theta)$ is a consistent estimator of the covariance of $\sqrt{n}\widehat{U}(\theta)$ such as a method of moments estimator. Additionally, $\widehat{\mathcal{I}}(\widehat{\theta})$ is Fisher's observed information matrix with sampling weights,

$$\widehat{\mathcal{I}}(\theta) = -\frac{\partial \widehat{U}(\theta)}{\partial \theta} = -\sum_{i=1}^{n} w_i \frac{\partial^2 \ell_i(\theta)}{\partial \theta \partial \theta^T}.$$

The KL divergence for the fitted model with the estimate $\widehat{\theta}$ is

$$KL(f, \widehat{f}_{\widehat{\theta}}) = E_f\left[\log\left\{\frac{f(x)}{\widehat{f}_\theta(x)}\right\}\right] = E_g\big[\log f(x)\big] - \ell(\widehat{\theta}).$$

To select the model with the smallest KL divergence we can select the model with the largest $\ell(\widehat{\theta})$. To calculate the AIC we need to estimate $E_g[\ell(\widehat{\theta})]$. A naive estimate of the log-likelihood is the log-likelihood fit to the training data $\widehat{\ell}(\widehat{\theta})$. This can be thought of as -1 times the weighted training error under negative log-likelihood loss. As in the unweighted case, the training error will likely underestimate the test error so $\widehat{\ell}(\widehat{\theta})$ should overestimate $E_g[\ell(\widehat{\theta})]$. Namely,

$$E_g[\widehat{\ell}(\widehat{\theta})] = E_g[\ell(\widehat{\theta})] + \frac{1}{n}\text{trace}(\Delta) + o_p(1/n)$$

where $\Delta = E[\widehat{\mathcal{I}}(\theta^*)]V(\theta^*)$. The proof is in the Appendix A.1 of [76]. Thus, the expected population log-likelihood $E_g[\ell(\widehat{\theta})]$ can be estimated by,

$$\widehat{\ell}(\widehat{\theta}) - p\bar{\delta}/n \tag{2.23}$$

where $\bar{\delta} = \text{trace}(\Delta)/p$ which can in turn be estimated by $\hat{\bar{\delta}} = \text{trace}(n\widehat{\mathcal{I}}(\widehat{\theta})\widehat{V}(\widehat{\theta}))/p$. Let loglik $= n\widehat{\ell}(\widehat{\theta}) = n\sum_{i=1}^{n} w_i\ell_i(\widehat{\theta})$ which is the sum over all observations since the weights sum to one. We then multiply this estimate by -2 to match the original AIC formula and obtain the design-based AIC,

$$d\text{AIC} = -2\text{loglik} + 2p\hat{\bar{\delta}}.$$

This extends the definition of AIC (Equation 2.19) by inflating the penalty term with the average estimated design effect $\hat{\bar{\delta}}$. If the weights are uniform and the true model is included in the sampling scope (so Δ is the identity matrix), then the dAIC reduces to the standard AIC.

Horvitz-Thompson-Efron estimator

Holbrook, Lumley and Gillen [57] combined the idea from Efron's general covariance penalty and the design-based AIC. The resulting estimator, named the Horvitz-Thompson-Efron (HTE) Estimator is an analytic estimate of the prediction error for the *q class* of error measures that accounts for the sampling scheme by including sampling weights. The HTE estimator is,

$$\text{HTE} = \sum_{i=1}^{n} w_i(\text{err}_i + 2\text{Cov}(\widehat{\lambda}_i, y_i)). \tag{2.24}$$

The *q class* of error measures and $\widehat{\lambda}_i$ are defined and discussed in Section 2.3.5. The covariance between $\widehat{\lambda}_i$ and y_i can be estimated with a parametric bootstrap if there is not an analytic form.

In the conclusion of their paper, Holbrook, Lumley, and Gillen suggested it would be useful to develop weighted resampling estimates of the prediction error for survey samples.

Weighted cross-validation

A recent paper by Wieczorek et al. (2022) proposed several analogues of K-fold cross-validation for survey samples [115]. A natural extension to the cross-validation estimator for the EPE under non-uniform sampling (Equation 2.20) is

$$\mathrm{CV}_{K,w} = \sum_{i=1}^{n} w_i L(y_i, \widehat{f}^{-k(i)}(x_i)) \tag{2.25}$$

where the weights w_i sum to one. Wieczorek et al. discussed how to construct the folds for this weighted K-fold cross-validation estimator of the EPE. First, they discussed unequal probablity sampling, such as the when units are sampled based on prespecified sampling probabilities. For this case, they proposed following the standard approach where each fold is a simple random sample from the larger training sample. They cite a Lemma from Cheng et al. 2010 [27] that states this approach will provide folds with sampling probabilities for each unit proportional to their sampling probability from the full training sample. They also discuss cross-validation for cluster samples and stratified samples. They suggest drawing the folds with cluster and stratified sampling, respectively, to mimic the sampling scheme of the original sample.

After the folds are determined they suggest following the standard cross-validation approach of mimicking the full analysis within each cluster regardless of if the original sampling scheme involved clustering, stratification, or unequal probability sampling. For each fold, the weighted prediction model should be fit on all folds besides the given fold and the weighted loss function is computed on that fold.

In a simulation study they found that selecting folds using cluster sampling or stratified sampling if the original sample was collected using cluster sampling or stratified sampling, respectively, was a better estimate of the true EPE than drawing the folds using a simple random sample. Additionally, when the sample was drawn with informative sampling, they found that weighting both the prediction model and the estimate of the EPE selected the correct model but omitting weights in one or both of these steps selected a less optimal model in their scenario.

Wieczorek et al. concluded that the weighted cross-validation estimates of the EPE out performed the unweighted cross-validation estimates. They did not, however, compare weighted resampling estimates to weighted analytic estimates.

2.3.9 Comparing Analytic and Resampling Methods Under Non-Uniform Sampling

Lumley and Scott derived a relationship between the design-based AIC and weighted leave-one-out cross-validation estimators [76]. They extended the relationship between AIC and the LOOCV under uniform sampling from Stone (1977) [107] discussed in Section 2.3.7. The weighted LOO CV estimator of the expected population log-likelihood is

$$\widehat{\ell}_{\text{LOO}} = \widehat{\ell}(\widehat{\theta}_{(i)}) = \sum_{i=1}^{n} w_i \ell_i(\widehat{\theta}_{(i)})$$

where $\widehat{\theta}_{(i)}$ is the vector of parameters from fitting the model without observation i. Lumley and Scott showed (see Section 3.1 and Appendix A.1 for the proof) that this can be related to the model fit on the full data as follows,

$$\widehat{\ell}_{\text{LOO}} = \widehat{\ell}(\widehat{\theta}) - \text{trace}\big(\widehat{\mathcal{I}}(\widehat{\theta})\widehat{V}_J\big) + o_p(n^{-1})$$

where $\widehat{V}_J$ is a jackknife estimator of $\text{Cov}(\widehat{\theta})$:

$$\widehat{V}_J = \frac{n-1}{n} \sum_{i=1}^{n} \big(\widehat{\theta}_{(i)} - \widehat{\theta}\big)^2.$$

From Equation 2.23, the dAIC estimator of the expected population log-likelihood is,

$$\widehat{\ell}_{d\text{AIC}} = \widehat{\ell}(\widehat{\theta}) - \text{trace}\big(\widehat{\mathcal{I}}\big(\widehat{\theta}\big)\widehat{V}(\widehat{\theta})\big).$$

Recall that $\widehat{V}(\widehat{\theta})$ (Equation 2.22) is a consistent estimator of the covariance of $\widehat{\theta}$. As long as $\widehat{V}_J$ is also a consistent estimator of the covariance, the dAIC and weighted LOO cross-validation estimators of the prediction error are asymptotically equivalent.

Although previous papers have evaluated the performance of weighted analytic estimates [76, 57] and weighted resampling methods [115] individually, to our knowledge no one has compared the methods to each other. Resampling methods should be a better estimate of the EPE because they directly estimate the quantity instead of targeting the in-sample error, but they are more computationally expensive. It would be useful to know how much better resampling methods perform and their computational burden compared to analytic methods with non-uniform sampling weights. Additionally, no one has evaluated if the dAIC and weighted leave-one-out cross validation estimators are asymptotically equivalent in a simulation study. We address these gaps in Chapter 5.

Chapter 3

Adjustment for Biased Sampling Using NHANES Derived Propensity Weights

3.1 Introduction

In this chapter, we discuss a method for estimating sampling weights for biased samples. As mentioned in the introduction, the Consent-to-Contact (C2C) registry at the University of California, Irvine (`https://c2c.uci.edu`) enrolls potential participants to aid in clinical research recruitment strategies [51]. Participants are recruited into the registry through a variety of outreach strategies including emails, community talks, postcards, and other methods. Due to this the C2C is not expected to be representative of the United States (US) population. For example, C2C participants tend to report more years of education relative to the general population, are more likely to be non-Hispanic White, have lower rates of comorbidities and higher rates of exercise (see Figure 3.3). C2C participants self-report demographic and clinical characteristics during the enrollment process. They also indicate their willingness to be contacted for potential participation in studies that involve various procedures or requirements such as lifestyle/behavioral modification, medication use, blood collection, brain imaging, autopsy, or lumbar punctures. Depending on the requirements and enrollment criteria of a study, participants who specifically report willingness to be contacted about required procedures are likely to be eligible and can be invited to participate, increasing the efficiency of recruitment.

Due to the under-representation of racial/ethnic minority populations in clinical research [87], members from our research team used data from the C2C to examine differences in willingness to participate by race/ethnicity [100]. Since the C2C is unrepresentative of the target population, using it directly could lead to potentially biased estimates and limited generalizability of estimated associations. For example, one natural way this bias could arise is through a differential relationship between race/ethnicity and willingness to participate by education level.

Convenience samples, like the C2C, are widely used to answer scientific questions because samples representative of the population may be impractical or unethical to collect. Most statistical methods assume representative sampling, but may be naively applied to biased samples. Participant self-selection into convenience samples may, however, reduce the degree to which such samples are representative of target populations of interest leading to a biased sample.

Some common approaches to address biased samples from self-selection, or selection bias, incorporate outside information about a population to obtain more generalizable inference. MRP (multilevel regression and poststratification) derives subpopulation estimates from national surveys. Iterative proportional fitting (or raking) adjusts subpopulation counts to match known marginal counts. For examples of MRP see [45] and [89], for an application of MRP see [102] and for raking see [17] and [71]. Additionally, there are several methods developed to combine estimates from multiple surveys, such as small-area estimation, joint-modeling, and imputation based methods. For a summary see [39].

Inverse probability of sampling weights is a popular solution for obtaining more generalizable inference and this approach is the focus of this chapter. Inverse probability weights are commonly used to adjust for design-based sampling of subpopulations and provide generalizable inference [75]. For design-based survey sampling, inverse probability weights are generally prespecified and fixed by design as the inverse of the sampling probability for each unit. Weights can also be used to account for selection bias in convenience samples, but in this context they are not fixed or known. Convenience samples often do not reflect the prespecified target population of interest because subjects self-select to participate. One solution is to use a representative sample to estimate sampling probabilities and the corresponding sampling weights for subjects in convenience samples (see for example [26], [2], [94], [40], [117], and [88]).

Estimating sampling weights for convenience samples can only be done if a relevant representative dataset is available. Most of the previous work on calibrating sampling weights assume that a representative sample or a sample frame was readily available (as in [117] and [88]). The National Health and Nutrition Examination Survey (NHANES) is a practical dataset for estimating propensity weights in biomedical convenience samples such as the C2C because it is representative of the US population, it contains medical measurements, and the data are open access (`https://wwwn.cdc.gov/nchs/nhanes/Default.aspx`). NHANES recruits approximately 5,000 individuals from across the US each year and over-samples people over 65 and minority groups, i.e. Hispanic, non-Hispanic (NH) Black and NH Asian subjects. NHANES data comes with sampling weights for each subject that are a function of the sampling probabilities and can be used to obtain a dataset that is representative of the US. Medical and dietary information are collected through structured questionnaires and in-person measurements [25]. It is common to compare estimates of population parameters, such as the prevalence of diabetes in the US, to estimates from NHANES as a diagnostic check for selection bias [43, 49, 11], but it is not often used for estimating sampling weights. Concurrent work from [2] proposed the use of national survey samples, including NHANES, as a practical way to obtain a representative sample to generalize randomized trial results. They used sampling weights for the complex survey sample in the propensity weight estimation model and the final outcome model, but we utilize them as frequency weights and generate a representative pseudopopulation.

The goal of this chapter is to obtain generalizable estimates of a scientific association from a convenience sample while correcting for sampling bias using a representative sample. In our setting, we are unable to obtain a pooled estimate across the convenience sample and representative sample because the outcome of scientific interest is not observed in the representative dataset. In this chapter, we focus on estimating inverse probability weights or, as we refer to them in this chapter, propensity weights for inclusion into a convenience sample. The propensity weights should not be confused with the propensity score discussed in Chapter 4. We do this by combining a convenience and representative sample and use the commonly collected covariates between the two to estimate the probability of membership in the convenience sample versus the representative sample. Our goal is to make this method easy to implement by using NHANES as a representative dataset and providing a package in *R*. We further derive an analytic variance estimator that extends the sandwich estimator for survey weighted generalized linear models [77] to account for uncertainty from estimating the propensity weights. We follow a similar approach to [103] and treat the sampling weight estimation model and final outcome model as being simultaneously estimated. Recent work by [113] and work by [26] also use a simultaneous estimation procedure for estimating the variance in the final parameter estimates when using estimated sampling weights. Our approach is similar, but we derive the variance of coefficient estimates in the final outcome model, instead of for population mean estimates. Alternatively, [2] used a double-bootstrap to account for uncertainty from estimating propensity weights and from the impact of non-response on the complex survey sample weights. We apply these methods to obtain valid population-level inference for the C2C registry.

The remainder of the manuscript is organized as follows: In Section 3.2, we develop the proposed methodology for calibrating propensity weights and quantifying uncertainty in the final scientific model of interest. In Section 3.3, we present a simulation study investigating the impact of estimated propensity weights on bias and variance. In Section 3.4, we apply our method to an analysis of racial and ethnic differences in research willingness. Finally, we conclude with a discussion of the advantages and limitations of the proposed method.

3.2 Methodology

Consider two collections of variables, $\mathcal{X}_R$ and $\mathcal{X}_C$ and let $\mathcal{X}$ be the set of variables in both $\mathcal{X}_R$ and $\mathcal{X}_C$, or $\mathcal{X} \equiv \mathcal{X}_R \cap \mathcal{X}_C$. Further, let $\mathcal{Y}$ be a subset of the variables in $\mathcal{X}_C$ that are not in $\mathcal{X}_R$, i.e., $\mathcal{Y} \in \mathcal{X}_C \setminus \mathcal{X}$. We will assume that it is possible to collect data sets on variables from both $\mathcal{X}_R$ and $\mathcal{X}_C$, but that it is not possible to collect random samples for all of the variables from $\mathcal{X}_C$ and thus difficult to obtain population inference. A data set obtained on $\mathcal{X}_C$ will be "convenience" and thus subject to potential bias (such as self-selection bias among other potential issues). Our ultimate goal is population-based inference on the set of variables in $\mathcal{Y}$ that are only collected in the convenience sample. For example, we may be interested in the estimation of the association between some subset of variables from $\mathcal{X}$, defined as $\mathcal{Z}$, i.e., $\mathcal{Z} \subset \mathcal{X}$, and $\mathcal{Y}$. In our context $\mathcal{Z}$ is race/ethnicity (available in both NHANES and C2C) and $\mathcal{Y}$ is willingness to participate in research (available only in C2C). This is not possible using $\mathcal{X}_C$ alone and so our goal is to leverage data sets collected through random sampling ($\mathcal{X}_R$) and convenience ($\mathcal{X}_C$) to obtain valid population-based inference for $\mathcal{Y}$.

To accomplish this goal, we employ weighted estimators for samples on variables from $\mathcal{X}_C$

that are constructed to obtain population-based inference. To do this, we estimate weights, w_C, that leverage information about the differences in the sample distributions between data sets on the variables from $\mathcal{X}_C$ and $\mathcal{X}_R$. Let X, Y, and Z be samples of observations on variables from the sets $\mathcal{X}$, $\mathcal{Y}$, and $\mathcal{Z}$ respectively. The general approach for estimating weights is to first collect data sets, X_C and X_R on variables from $\mathcal{X}_C$ and $\mathcal{X}_R$, respectively, and then estimate a model to discriminate between the two datasets. To accomplish this we stack X_C onto X_R to a create a single combined dataset. We then construct an auxiliary variable, C, which is an indicator that an observation was in dataset X_C. Thus, C evaluates to 1 for units from X_C and 0 otherwise. We can use this stacked data set to estimate the probability of a given unit arising from each sample and use this information to weight X_C to be similar to X_R and obtain inference on variables from $\mathcal{Y}$.

Let the subscript i denote the observation for subject i. Our goal is to obtain a weighted estimator for population-based inference such that,

$$E_{\mathcal{P}|C}[w_{Ci} Y_i | C_i = 1] = E_{\mathcal{P}}[Y_i] \tag{3.1}$$

where $E_{\mathcal{P}}$ is the expectation taken over the distribution of the target population, $\mathcal{P}$. Now, we define,

$$P_{Ci} = \Pr(C_i = 1 | X_i = x_i), \tag{3.2}$$

which allows us to construct weights

$$w_{Ci} \propto \frac{1 - P_{Ci}}{P_{Ci}} \tag{3.3}$$

so that Equation 3.1 holds under appropriate assumptions on the set $\mathcal{X}$ (see Section 3.2.1). Weights will be normalized so they sum to one. Equation 3.2 is analogous to the propensity score from the causal inference literature [95] and Equation 3.3 would correspond to the weights for the average treatment effect on the control (ATC) [69]. Thus, we refer to the estimated weights w_C as inverse propensity weights for self-selection into the convenience sample, or propensity weights for short.

3.2.1 Assumptions

While our focus throughout will be obtaining population inference from a convenience sample, there are many analogs between the methods here (and their assumptions) and those from the field of causal inference. See Section 2.2 for more details about the assumptions. Therefore, we will describe the assumptions needed for estimating population parameters through the lens of estimating causal effects and discuss how these assumptions do, and do not, apply in our context. The methods here relate most closely to those involving the "propensity score" [95].

A propensity score as defined by Rosenbaum and Rubin [95] is the probability of receiving a treatment when conditioned on observed covariates. In the field of causal inference, typically three assumptions are needed for making causal conclusions using propensity scores: (1) unconfoundedness, (2) positivity, and (3) the stable unit treatment value assumption (SUTVA) (See Chapter 1 and 12 of [62] or Appendix 1 of [50] for an overview). First, the (1) unconfoundedness assumption requires that potential outcomes be conditionally independent of the treatment assignment given the observed covariates. Second, (2) positivity requires that each unit has a positive probability of receiving both treatment and control treatments. More formally, if T is an indicator for receiving a treatment and X are covariates, then $0 < \Pr(T = 1|X = x) < 1$ for all subjects. Finally, (3) SUTVA requires that each subjects treatment assignment does not affect any other subject's potential outcomes and there is no hidden variability in the treatment.

Versions of two of these assumptions are relevant in our context with propensity weights.

First, we assume (1) unconfoundedness, that the response is independent of the selection probability conditional on the collected covariates. In other words, any covariate (or a proxy of the covariate) related to both the response or selection probability must be measured in both the representative and biased samples. We are unable to balance on unmeasured covariates. For (2) we assume that each subject must have a non-zero probability of being selected into the convenience sample, or $0 < \Pr(C = 1|X = x) < 1$. Although C2C registration is open to any adult, participants are primarily enrolled from Southern California. Although theoretically possible, the probability of people from outside of the Southern California region being sampled in the C2C is close to zero. Thus, we must assume that the relationship between race/ethnicity and research willingness does not vary by state within the US.

The SUTVA assumption (3) has less applicability in our setting. In particular, SUTVA is typically used to make a consistency argument to map potential outcomes to observed outcomes. In our setting, we assume that the response of an individual would be the same whether or not they are in the convenience sample or the random sample, and therefore the need for this assumption in our setting is diminished. More explicitly, we assume no effect of survey participation on the potential outcome of a participant and our propensity weights are being used to reweight participants back to their population prevalence based on the rates from the random sample.

In practice, it is important to carefully design the data collection to include any covariates hypothesized to be related to the sampling probability. If there are any missing covariates, accounting for the sampling bias due to measured covariates should be better than ignoring the selection mechanism completely [80], but the unconfoundedness assumption is not testable [62]. The non-zero sampling probability assumption should motivate researchers to collect participants from each subpopulation based on variables related to selection. We can upweight underrepresented subpopulations, but we are never able to learn about subpopulations that were never studied.

3.2.2 Dataset Construction for Estimating Propensity Weights

We want to estimate propensity weights for the convenience sample X_C with n_C observations. Let m and p be the number of variables in $\mathcal{X}$ and $\mathcal{Z}$, respectively. We include a column of 1s in the sets $\mathcal{X}$ and $\mathcal{Z}$ to be able to estimate an intercept. When using a complex survey sample, such as NHANES, as the representative sample we need to first incorporate design weights to ensure it is representative of the population of interest because certain subpopulations may be oversampled by design. Let X_S be a survey sample with n_S observations and P_{Si} denote the sampling probability for subject i in the survey sample. In NHANES, the sampling probabilities for each subject account for both the survey design and both item and subject level non-response. To obtain a representative dataset, we utilize frequency weights, $w_{si} = P_{Si}^{-1}$, which represent the number of subjects each sampled subject represents in the population and replicate each subject according to their frequency weight. To obtain the smallest representative dataset with whole numbers of subjects, we divide each frequency weight by the smallest observed weight and take the ceiling of it to obtain the number of replications: $w^*_{Si} = \text{Ceiling}(w_{Si}/\text{min}[w_S])$. We implement the frequency weights to obtain a representative sample X_R with dimension $n_R \times m$, where each subject from X_S is replicated w^*_{Si} times for a total of $n_R = \sum_{i=1}^{n_S} w^*_{Si}$ observations. While this approach does not fully account for clustering in the NHANES dataset because information on clustering is not publically available, it is a pragmatic solution because the constructed sample will be more representative than most convenience samples thereby leading to improved inference on the target population. Additionally, clustering may slightly impact the variance of parameters, but in this project, we do not account for variability in the NHANES sampling weights.

Recall that X is the sample of variables collected in both X_R and X_C and containing observations from both representative and convenience sample subjects. Specifically, X is an $n \times m$ dimension matrix where $n = n_R + n_C$. For notational convenience, let $\mathcal{C}$ be the set of subjects from the convenience sample with $|\mathcal{C}| = n_C$ and and $\mathcal{R}$ be the set of subjects from the representative sample with $|\mathcal{R}| = n_R$. To obtain X in practice, we concatenate the convenience and representative samples for the variables in $\mathcal{X}$. We can derive the indicator for membership in the convenience sample, C, and append it to X. To estimate the propensity weights, w_C, defined in Equation 3.3 we can directly estimate the probability of convenience sample membership, P_{Ci}, defined in Equation 3.2. Many types of propensity weight estimation methods can be considered and we discuss their advantages and disadvantages in the following section.

3.2.3 Classes of Propensity Weight Estimation Methods

In this section we compare different sampling weight estimation strategies and provide examples of each that we will use as test cases. Our goal is to assess the relative performance of different estimation strategies and their strengths and weaknesses. We discuss likelihood based methods and use logistic regression as an example, covariate balancing methods with the covariate balancing propensity score and entropy balancing as examples, and algorithmic methods with random forest as an example. We will explore these four examples of methods for estimating propensity weights and their impact on covariate balance of convenience samples and on bias and variance of estimated associations.

Likelihood-based methods such as linear regression, logistic regression, probit regression, and penalized regression minimize the negative log-likelihood. We focus on logistic regression which takes the form $\text{logit}(P_{Ci}) = X_i\gamma$, where $\text{logit}(\cdot)$ is the logit function, γ is a $m \times 1$ vector of regression coefficients and X_i is the $m \times 1$ vector of observed covariates for subject i. When implementing logistic regression, we include second order terms and use forward-selection with Akaike's Information Criterion (AIC) for selecting predictors using the `step` function in the `stats` package.

By definition, propensity scores are balancing scores [95] and another strategy for propensity weight estimation is to directly balance covariate distributions across the two classes of a binary outcome. Covariate balancing methods, such as the covariate balancing propensity score (CBPS) and entropy balancing (EB), optimize weights by directly targeting covariate balance between the the convenience sample ($C_i = 1$) and the representative sample ($C_i = 0$).

CBPS extends the logistic regression model by incorporating additional moment balancing constraints [61]. CBPS is a common method used for estimating propensity scores (see for example [116], [44], and [82]). Researchers may be familiar with CBPS and interested in using it for estimating propensity weights, so we will assess how it performs relative to logistic regression. In our context, estimated propensity weights should balance the covariate distribution of the convenience sample so it matches the representative sample [69]. Thus, we estimate weights for the average treatment effect on the control (ATC) so that the representative sample is the reference population. The CBPS method solves the estimating equations as well as covariate balancing conditions. For the ATC, the balancing conditions are

$$E_{\mathcal{P}}\left\{\frac{(1-P_{Ci})C_i f(X_{ij})}{P_{Ci}} - (1-C_i)f(X_{ij})\right\} = 0, \tag{3.4}$$

where $f(X_{ij})$ is a function of X_{ij}. For example, $f(X_{ij}) = (X_{ij}; X_{ij}^2)$ would ensure the first and second moments of each covariate will be balanced. We fit the CBPS model with the `CBPS` package and included balancing constraints on second order orthogonal polynomial terms.

Entropy balancing (EB) is a non-parametric approach to weight estimation that incorporates moment balancing conditions into model selection [52]. Entropy balancing allows for the inclusion of initial base weights that contain information about population prevalence. Entropy balancing estimates weights that minimize the divergence between estimated weights and base weights. If there are no base weights available, they can be treated as uniform for all units ($b_i = 1/n$). Similar to the CBPS, we want to estimate weights for the ATC where the representative sample is considered the reference group. Assuming Kullback–Leibler divergence, EB minimizes $H(w_C) = \sum_{i \in \mathcal{C}} w_{Ci} \log(w_{Ci}/b_i)$, subject to balancing constraints,

$$\frac{1}{n_R} \sum_{i \in \mathcal{R}} x_{ij}^d = \sum_{i \in \mathcal{C}} w_{Ci} x_{ij}^d, \text{ for } d = 1, ..., D,$$

where x_{ij} is the j-th covariate for subject i, $\sum w_{Ci} = 1$, and $w_{Ci} > 0$. We implemented entropy balancing with $D = 3$ using the `entbal` package for `R` (`https://github.com/bvegetabile/entbal`) available on GitHub [110].

Unconstrained algorithmic methods, such as support-vector machines or random forest, do not require specifying a model. They predict class probabilities that minimize some out-of-sample measure of goodness of fit (e.g. prediction error) and do not assume a distribution or any moments of the response. As an example, we focus on random forest [19] which is a flexible model that does not require the user to specify a functional form of predictors. Random forest (RF) is an extension of classification and regression trees or CART [20] and limits susceptibility to overfitting by introducing stochasticity. RF builds separate decision trees on bootstrapped samples and averages the prediction across trees for each subject, a technique known as bagging, and for each node of a decision tree, only a random sample of predictors are considered for splitting. To prevent extreme weight values, we trim RF estimates of the probability of convenience sample membership that are 0 or 1 and replace them estimate with 0.01 and 0.99, respectively [67]. We fit RF using the `randomForest` package in `R` [70].

To demonstrate the advantages and disadvantages of the different approaches, we will evaluate the performance of the different propensity weight estimation methods described above: logistic regression, covariate balancing propensity score (CBPS), entropy balancing (EB), and random forest (RF).

3.2.4 Quantifying Uncertainty

There are several common ways to estimate the variance of coefficients from models with weights. Standard analytic variance estimates for coefficient estimates from weighted GLMs are an extension of the sandwich, or Huber-White, estimator [41]. These design based variance estimates are included in most survey sampling software packages such as the `svyglm` function in the `survey` package [77]. This approach assumes the propensity weights are fixed, i.e. not estimated, but uncertainty from propensity weight estimation will likely impact the uncertainty of the parameter estimates in the scientific model. If there is little uncertainty in $\widehat{w}_C$, then the design based errors used in survey sampling methodology will likely perform well. Alternatively, resampling methods, such as the bootstrap, are more computationally intensive, but can be used to account for the impact of the uncertainty in the propensity weight estimation procedure on the variance of the parameter estimates by reestimating weights within each bootstrap sample.

In this section, we derive an analytic variance estimate that accounts for uncertainty from the propensity weight estimation method. We use a simultaneous estimating equation approach for variance estimation and extend the approach of [103]. We treat both the propensity weight estimation and final scientific model as if they are being estimated simultaneously and derive the sandwich estimator for the parameters of the scientific model. Of the four propensity weight estimation methods we have considered, only the logistic model allows for a readily tractable analytic variance estimate that accounts for uncertainty in propensity weights. The design based variance estimate can be used for other weight estimation methods.

Suppose the scientific outcome model with response, Y_i, and the $p \times 1$ vector of covariates, Z_i, for subject i is $\eta_i = g(\mu_i) = z_i\beta$ where η_i is the linear predictor, $g(\cdot)$ is a link function, $\mu_i = E(Y_i|Z_i = z_i)$, and β is a $p \times 1$ vector of parameters. Recall, when applying this method to the analysis of the C2C, Z is the subset of covariates in the C2C sample needed for the final analysis, which includes race/ethnicity and adjustment variables, and Y is the willingness to participate. Assuming a representative sampling scheme, the i-th observation's contributions to the k-th element of the score equation is given by

$$U_{ki}(\beta) = \Big(\frac{\partial \mu_i}{\partial \beta_k}\Big)\Big(\frac{Y_i - \mu_i}{V(\mu_i)}\Big)$$

where $V(\mu_i) = Var(Y_i)$ and for $k = 1...p$ [85]. When using an unrepresentative sample, each subjects contribution to the score is weighted by their propensity weight, w_{Ci} as follows,

$$\overline{U}_k(\beta) = \sum_{i \in \mathcal{C}} w_{Ci} U_{ki}(\beta) = 0.$$

This notation deviates from the notation introduced in Section 2.3.8 and instead follows the notation from [77]. Define the logistic regression propensity weight model, ψ_i, as $\psi_i = \text{logit}(P_{Ci}) = x_i\gamma$ where γ is a $m \times 1$ vector of coefficients. Let x_i be the i-th row of X, the combined $n \times m$ matrix including covariates from the convenience sample and representative sample. The estimated propensity weights w_{Ci} are a function of the probability of convenience sample membership (Equation 3.3) and sum to one. The $m \times 1$ score equation has element j,

$$T_j(\gamma) = \sum_{i \in \mathcal{C} \cup \mathcal{R}} T_{ji}(\gamma) = \sum_{i \in \mathcal{C} \cup \mathcal{R}} (C_i - P_{Ci}) x_{ij} = 0, \tag{3.5}$$

where $j = 1...m$. We consider both the score equation for the propensity weight estimation, $T_i(\gamma)$, and the score equation for the scientific outcome model, $\overline{U}_i(\beta, \gamma) = \overline{U}_i(\beta)$. We include γ in the notation to emphasize that the score for the final scientific outcome is a function of the γ through the propensity weights as in [103]. To simplify notation, we sometimes refer to $T_i(\gamma)$ and $\overline{U}_i(\beta, \gamma)$ as T_i and $\overline{U}_i$, respectively. We combine the two estimation equations into a stacked estimating equation

$$\begin{pmatrix} \sum_{i\in\mathcal{C}\cup\mathcal{R}} T_i(\gamma) \\ \\ \sum_{i\in\mathcal{C}} \overline{U}_i(\beta,\gamma) \end{pmatrix} = 0. \tag{3.6}$$

Using a first order Taylor series expansion of the stacked estimating equation (Equation 3.6) we obtain the variance estimate,

$$\widehat{V}_{Prop}[(\widehat{\gamma},\widehat{\beta})] = \widehat{I}^{-1}\widehat{Q}\widehat{I}^{-1}, \tag{3.7}$$

where the parameters have been replaced with maximum likelihood estimates. In Equation 3.7, I is Fisher's information matrix under the assumed distributions of Y_i and C_i such that,

$$I = \begin{pmatrix} I_{TT} & 0 \\ I_{UT} & I_{UU} \end{pmatrix}$$

and Q is the true variance of the score, where

$$Q = \mathrm{Var}\left(\begin{matrix} \sum_{i\in\mathcal{C}\cup\mathcal{R}} T_i(\gamma) \\ \\ \sum_{i\in\mathcal{C}} \overline{U}_i(\beta,\gamma) \end{matrix} \middle| X = x, Z = z \right) = \begin{pmatrix} E_{\mathcal{P}}[TT^T|X=x] & R^T \\ \\ R & E_{\mathcal{P}}[\overline{U}\,\overline{U}^T|Z=z] \end{pmatrix}.$$

Derivations of the components of I and Q are provided in Appendix A. The reader may notice similarities to the design based variance used in the survey sampling literature without the finite population correction factor [77]. Define $\widehat{A} = \widehat{I}_{UU}$ and $\widehat{B} = \overline{UU}^T$ so the proposed variance estimator is,

$$\widehat{V}_{Prop}(\widehat{\beta}) = \widehat{A}^{-1}\widehat{B}\widehat{A}^{-1} - \widehat{A}^{-1}\widehat{I}_{UT}\widehat{I}_{TT}^{-1}\widehat{R}^T\widehat{A}^{-1}.$$

Thus the proposed variance estimate can be expressed as the standard design based variance estimator plus a correction factor. We have provided a `estweight` package available for `R` on GitHub (`https://github.com/oliviabern/estweight`). The `convGLM` function takes a representative sample, convenience sample, and the final outcome model and provides weighted parameter estimates. If the user selects a logistic propensity weight estimation method, the function returns the proposed variance estimate, otherwise it provides standard design-based variance estimates.

3.3 Simulation Studies

We considered the impact of our proposed weight estimation on bias of estimated associations and the accuracy of uncertainty quantification procedures through empirical simulation studies. We designed a simulation study to be similar to the analysis of Salazar et al. (2020) [100]. NHANES collects cross-sectional data on a 2-year cycle so we combined data from the 2013-2014 and 2015-2016 surveys. All simulations utilized the NHANES data, and like Salazar et al., we excluded all subjects with a reported race or ethnicity of "other" for a total of $n_S = 4,471$ subjects. All subjects had complete data on age, sex, education, race, ethnicity, medical history (high blood pressure, diabetes, kidney disease, liver disease, coronary heart disease, cancer, major depression, prescription drug use), exercise, and amount of sleep. To obtain a representative sample we replicated each observation according to their frequency weight for a final sample size of $n_R = 38,811$ observations. We refer to this representative dataset as NHANES-REP. Code for creating NHANES-REP and for reproducing the simulation study is available on GitHub (`https://github.com/oliviabern/estweight_simulationstudy`).

3.3.1 Simulation Set Up

To investigate the potential impact of underrepresentation in samples and estimated propensity weights on bias and variance of estimated associations we used NHANES-REP as a finite population and drew both representative and deliberately biased samples. Subjects who are Hispanic, NH Black, NH Asian, or who have lower education levels and do not exercise tend to be underrepresented in the C2C and so we generated smaller sampling probabilities for these subpopulations. We use $\mathbb{1}$ to denote an indicator variable. Let P_{Ci} be the biased sampling probability for subject i where $\text{logit}(P_{Ci}) = \psi_i$ with

$$\begin{aligned}\psi_i = &.15\mathbb{1}_{Female,i} + .25\mathbb{1}_{HighSchool,i} + .1\mathbb{1}_{<HighSchool,i} + .4\mathbb{1}_{SomeCollege,i}\\ &+ .85\mathbb{1}_{Hispanic,i} + .45\mathbb{1}_{NHAsian,i}\mathbb{1}_{NHAsian,i}\mathbb{1}_{SomeCollege,i}\\ &+ .05\mathbb{1}_{NHBlack,i} + .75\mathbb{1}_{NHBlack,i}\mathbb{1}_{Exercise,i} - .001Age_i^2 + 4.\end{aligned}$$

Within each simulation, we drew a representative simple random sample of size 500 and a biased sample of size 500 with sampling probabilities P_{Ci}. We simulated $Y_i \sim \text{Bernoulli}(\mu_i)$ with $\text{logit}(\mu_i) = \eta_i$ and

$$\begin{aligned}\eta_i = &1 + log(2)\mathbb{1}_{Hispanic,i} - log(3)\mathbb{1}_{NHAsian,i} + log(1.5)\mathbb{1}_{NHBlack,i} - log(2)P_{Ci}\\ &+ log(2)\mathbb{1}_{Hispanic,i}P_{Ci} + log(4)\mathbb{1}_{NHAsian,i}P_{Ci} - log(3)\mathbb{1}_{NHBlack,i}P_{Ci}.\end{aligned}$$

We estimated propensity weights for subjects in the biased sample with each of the four propensity weight estimation methods described in Section 3.2.3. Similar to our applied example, we were interested in a model of the the marginal relationship between race/ethnicity where,

$$\text{logit}(\Pr[Y_i = 1|X_i = x_i]) = \beta_0 + \beta_1\mathbb{1}_{Hispanic,i} + \beta_2\mathbb{1}_{NHAsian,i} + \beta_3\mathbb{1}_{NHBlack,i}.$$

For each simulation, we (1) fit the above model in the representative sample with the objective of obtaining a similar estimate using a biased sample. (2) We then fit the model in the biased sample without any weighting, (3) with the true propensity weights ($w_{Ci} \propto P_{Ci}^{-1}(1 - P_{Ci})$), and (4) with the estimated propensity weights from each of the four estimation methods and compared the estimates to those obtained in the representative sample. For CBPS and EB, we balanced continuous variables on the first and second moments such that $f(X_{ij}) = (X_{ij}, X_{ij}^2)$ for CBPS and $d = 2$ for EB. For the logistic regression model, we included second-order terms in the model scope and used stepwise AIC for variable selection. Note that we did not include interactions in the model scope so even though the data generating model is logistic, we were unable to correctly specify it. We computed and compared analytic and bootstrap estimates of the standard error to the empirical Monte Carlo standard error. To prevent under-representation of small subpopulations in bootstrap samples, we used race/ethnicity as a stratification variable for sampling. When stratifying the bootstrap sample failed to provide adequate representation of subpopulations leading to extreme weights and inestimable coefficients, we removed the parameter estimates from bootstrap variance estimates. We conducted 1,000 simulations and used 200 bootstrap samples within each simulation.

3.3.2 Simulation Results

Coefficient estimates

The average log odds ratios for representative and biased samples for each of the propensity weight types are shown in Figure 3.1. The goal of incorporating estimated weights is to match estimates fit using a biased sample (columns 2-7 of the tabulated results) to those estimated using a representative sample (column 1). Note that estimates derived from a biased sample that fail to account for the sampling scheme (i.e. no weighting, column 2) did

Mean Estimated Log Odds Ratios

	SRS	Biased Sample					
		None	True	Log.	CBPS	EB	RF
Intercept	0.64 (0.11)	0.56 (0.11)	0.65 (0.14)	0.64 (0.14)	0.63 (0.13)	0.64 (0.15)	0.67 (0.22)
Hispanic	1.09 (0.42)	1.19 (0.36)	1.10 (0.39)	1.12 (0.43)	1.12 (0.42)	1.12 (0.45)	1.09 (0.51)
NH Asian	−0.30 (0.50)	−0.17 (0.46)	−0.30 (0.52)	−0.29 (0.55)	−0.28 (0.53)	−0.29 (0.57)	−0.30 (0.63)
NH Black	−0.41 (0.31)	−0.48 (0.28)	−0.41 (0.33)	−0.39 (0.34)	−0.40 (0.33)	−0.39 (0.36)	−0.45 (0.43)

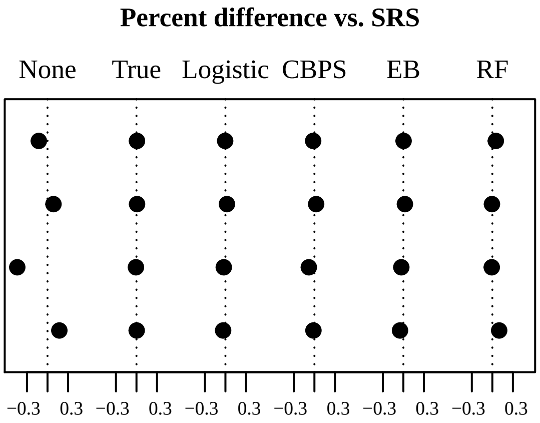


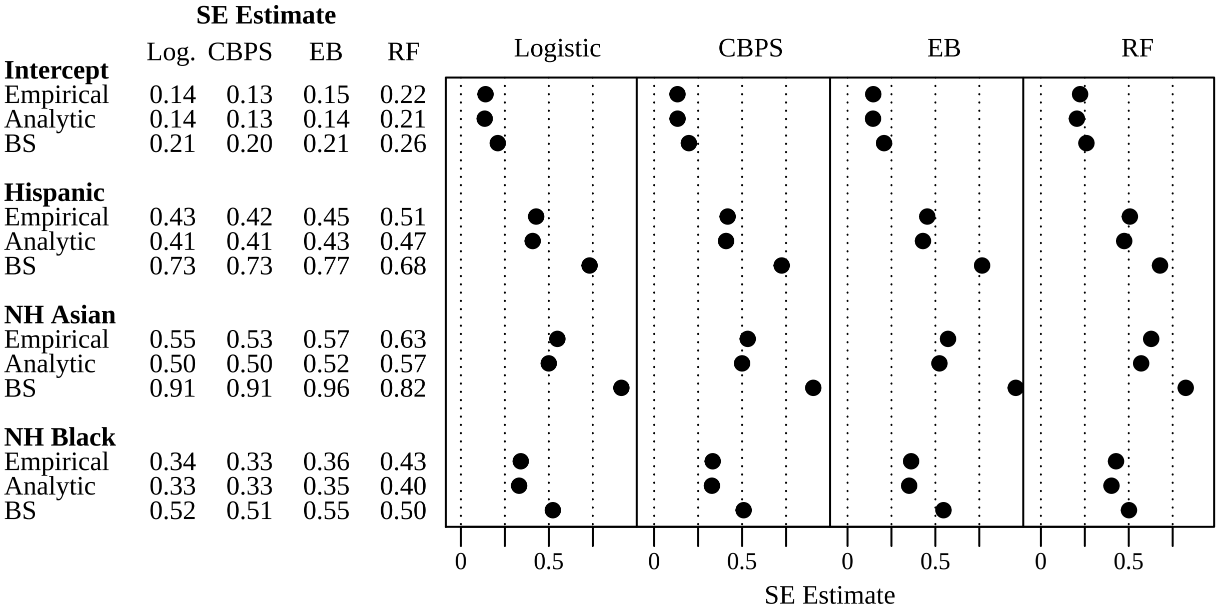
Figure 3.1: Results for the simulation study described in Section 3.3.1. The average estimated log odds ratios (empirical standard errors) are presented in the table on the left. Estimates for the marginalized model fit in the simple representative sample (SRS) are in column 1 and estimates fit in a biased sample along with different types of propensity weights are in the other columns. Results compare models fit in a biased sample that do not include weights (None), incorporate the true propensity weights (True), or incorporate propensity weights estimated with a logistic (Log.), covariate balancing propensity score (CBPS), entropy balancing (EB), or random forest (RF) approach. Percent bias comparing average estimates fit in a biased sample to estimates fit in a simple random sample (SRS) are presented in the figure on the right.

not match those from the representative sample. Incorporating both true (column 3) and estimated (columns 4-7) propensity weights allowed us to match the representative sample estimates. The type of propensity weight model did not have an appreciable impact as all of the weighted estimates did not vary much in comparison to the representative sample and to each other. Logistic regression performed well and allowed us to obtain weighted estimates fit in a biased samples that matched the association in the target population. It is a practical choice because it is parsimonious and easy to implement. It is important to note that the true sampling probabilities were generated from a logistic model, but the model included interactions that were not in the scope of the logistic model we used for estimation. In this example, incorporating estimated propensity weights was an effective method for obtaining inference on the target population.

Uncertainty estimates

In this simulation, empirical standard errors were larger for weighted estimates fit in biased samples compared to unweighted estimated fit in a representative sample (Figure 3.1). We also investigated the impact of the propensity weight estimation method on uncertainty and the performance of analytic and bootstrap variance estimates. Standard error estimates for the four different propensity weight estimation methods are reported in Figure 3.2. For the analytic variance estimate, we used the proposed analytic standard error estimate ($\widehat{V}_{Prop}$) when using weights estimated via logistic regression. When using weights estimated with CBPS, EB, and RF methods we used the design based errors from the `survey package` that ignore the propensity weight estimation [77]. The average analytic standard error estimate was generally comparable to the empirical standard error across simulations, but the bootstrap generally overestimated the true uncertainty. The design based and proposed variance estimates resulted in similar standard error estimates. Compared to the bootstrap estimate, analytic estimates more closely approximated the empirical standard error even though they did not account for uncertainty from variable selection when estimating propensity weights. Additionally, they were computationally easier than the bootstrap.

In 21 out of 1,000 simulations, coefficients were not estimable in some bootstrap samples due to extreme values of estimated weights. There was one bootstrap sample in two simulations where the association was inestimable when weights were estimated using EB. This was slightly more common when weights were estimated using RF–out of 1,000 simulations, 19 simulations had at most 5 bootstrap sample that were unable to estimate the association. We hypothesize that this occurs due to uniquely sparse bootstrap samples with little to no representation of some subpopulations, but this does not occur for logistic regression or CBPS. The RF method draws bootstrap samples to fit each tree and this bootstrap within a bootstrap may lead to subpopulations without any variation in the response, and thus extreme weights. Entropy balancing targets covariate balance which can be difficult if



SE Estimate	Log.	CBPS	EB	RF
Intercept				
Empirical	0.14	0.13	0.15	0.22
Analytic	0.14	0.13	0.14	0.21
BS	0.21	0.20	0.21	0.26
Hispanic				
Empirical	0.43	0.42	0.45	0.51
Analytic	0.41	0.41	0.43	0.47
BS	0.73	0.73	0.77	0.68
NH Asian				
Empirical	0.55	0.53	0.57	0.63
Analytic	0.50	0.50	0.52	0.57
BS	0.91	0.91	0.96	0.82
NH Black				
Empirical	0.34	0.33	0.36	0.43
Analytic	0.33	0.33	0.35	0.40
BS	0.52	0.51	0.55	0.50

Figure 3.2: Standard error estimates for the simulation study described in Section 3.3.1. The different standard error estimates (empirical, analytic and bootstrap) are reported for the coefficient estimates in the marginalized model fit in a biased sample. Standard error estimates are reported for models implementing propensity weights estimated using a logistic (Log.), covariate balancing propensity score (CBPS), entropy balancing (EB) or random forest (RF) method. Details of the standard error calculations are reported in Section 3.2.4

certain subpopulations are only observed in either the representative or convenience sample. Although CBPS also targets covariate balance, we did not observe any extreme weights and we hypothesize the focus on maximizing the likelihood may prevent this. Although this occurs infrequently, it may arise in practice. We excluded any bootstrap samples with insufficient information for estimating the association for a given sampling weight estimation method from the bootstrapped estimate of the variance.

3.4 Application to the C2C Willingness Analysis

The work of [100] investigated differences in research willingness by race and ethnicity using logistic regression. The methods are described in detail in that chapter and we summarize them here. They performed logistic regression models to assess racial/ethnic differences in willingness to participate in research in participants 50 years or older. They separately evaluated 9 different responses: willingness to be contacted about studies involving (1) physical activity/diet modification, (2) cognitive testing, (3) magnetic resonance imaging (MRI), (4) positron emission tomography (PET) scans, (5) blood draws, (6) approved medications, (7) investigational medications, (8) lumbar punctures and (9) autopsy. They adjusted for age, sex, educational attainment, number of comorbidities, number of medications, cognitive function instrument score [7, 112] and research attitudes questionnaire score [99] and used multiple imputation to handle missing data. C2C data is updated as more participants enroll and can be requested at `https://c2c.uci.edu/request-c2c-data/`. When replicating this analysis, we started with the same dataset of C2C participants, but performed our own multiple imputation.

3.4.1 Identifying Matching Covariates

We first identified covariates likely to modify the relationship between race/ethnicity and willingness to participate in research and were collected in both the C2C and NHANES. Some covariates were recorded with differing levels of granularity in the two datasets so we collapsed them into comparable subgroups. For example, the question regarding exercise was phrased differently in the two datasets. NHANES participants were asked if they participated in vigorous or moderate recreational activities in a typical week for at least 10 minutes.The C2C participants were asked if they participated in the following activities for at least 15 minutes/day at least once/week for the last year: walking, hiking, biking, aerobics, calisthenics, swimming, water aerobics, weight training, stretching, or another form of exercise. We decided to exclude the question about walking for the C2C subjects because there were high agreement rates for this question and we were concerned participants may have reported walking for purposes other than recreational exercise. To compare across groups, we created an indicator for exercise. We used NHANES participants who were 50 years or older to match inclusion criteria from [100]. In total, we included 14 variables: age, sex, education level [Educ.] (less than 12 years [< 12], high school/GED [12], some college [12-16], college graduate [16]), race/ethnicity (NH White, Hispanic, NH Asian, NH Black), high blood pressure (BP), kidney disease, liver disease, congestive heart failure (CHD), past cancer diagnosis, major depression, average hours of sleep per night, prescription medicine use (Presc. meds), and exercise.

Excluding subjects with a reported race or ethnicity of "other" resulted in $n_C = 2,749$ observations in the C2C and excluding 917 NHANES participants with missing data (out of $n_S = 5,605$) resulted in $n_R = 38,811$ observations in NHANES-REP. Most of the missingness in the C2C was confined to a few covariates: 576 subjects were missing sleep, 207 were missing prescription drugs, 167 were missing exercise, and 29 were missing history of cancer. Matching covariates for each dataset are summarized in Figure 3.3. Continuous covariates were summarized by mean (standard deviation) and the proportion was reported for categorical variables. Weighted sample statistics using estimated propensity weights for the C2C dataset were also presented. The propensity weights were estimated using logistic regression, CBPS, EB, and RF with one imputed C2C dataset. Logistic regression, CBPS, and EB balance the covariate distributions well, but RF weighted estimates are similar to unweighted ones.

3.4.2 Estimating Propensity Weights

We repeated the analysis performed by Salazar et al. and imputed 5 C2C datasets and used Rubin's rules to aggregate across datasets [98]. Within each dataset we estimated propensity weights for each subject using logistic regression, CBPS, EB, and RF. We fit the outcome models that quantify the relationship between race/ethnicity and each of the 9 outcomes and incorporated the estimated propensity weights. We report the estimated odds ratios (OR)

	NHANES −REP	C2C				
		None	Log.	CBPS	EB	RF
Age	63.6 (9.3)	65.9 (8.7)	62.4 (9.2)	62.5 (9.2)	63.6 (9.3)	64.2 (9.7)
Sleep (Hr)	7.4 (1.4)	6.7 (1.5)	7.2 (1.2)	7.2 (1.2)	7.5 (1.4)	7.2 (1.0)
Educ: < 12	0.16	0.01	0.12	0.12	0.16	0.02
Educ: 12	0.22	0.07	0.23	0.23	0.22	0.10
Educ: 12−16	0.31	0.19	0.31	0.31	0.31	0.26
Educ: 16	0.31	0.73	0.34	0.34	0.31	0.62
NH White	0.75	0.87	0.74	0.75	0.75	0.88
Hispanic	0.11	0.07	0.13	0.13	0.11	0.06
NH Asian	0.04	0.05	0.04	0.04	0.04	0.04
NH Black	0.11	0.01	0.08	0.08	0.11	0.02
Female	0.53	0.62	0.49	0.49	0.53	0.59
High BP	0.52	0.34	0.50	0.50	0.52	0.37
Diabetes	0.19	0.10	0.17	0.17	0.19	0.06
Kidney	0.05	0.03	0.04	0.04	0.05	0.01
Liver	0.05	0.02	0.06	0.06	0.05	0.01
CHD	0.09	0.06	0.09	0.09	0.09	0.03
Cancer	0.20	0.32	0.24	0.24	0.21	0.25
Exercise	0.46	0.83	0.51	0.51	0.48	0.75
Depression	0.10	0.12	0.07	0.07	0.10	0.04
Presc. meds	0.79	0.78	0.75	0.76	0.78	0.78

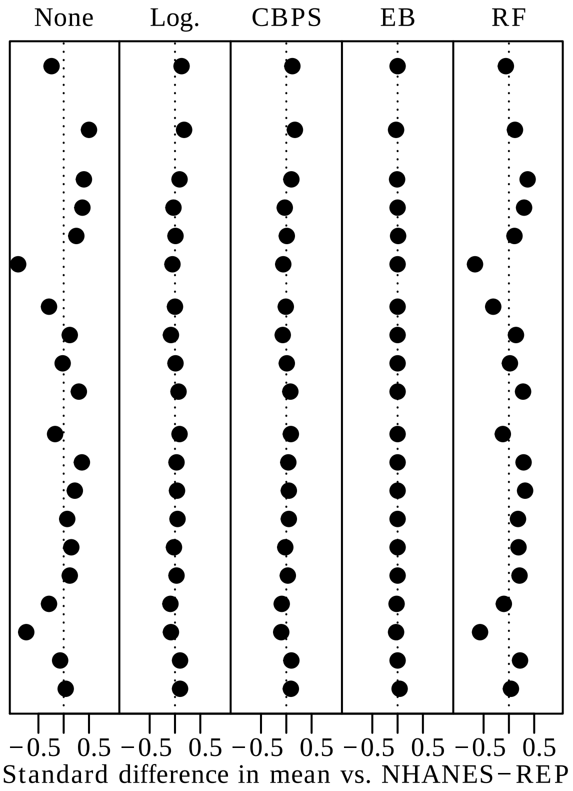


Figure 3.3: Covariates were summarized as mean (standard deviation) for continuous variables and as proportions for categorical variables in the table on the left for NHANES-REP (38,811 observations), the unweighted C2C (2,749 subjects) dataset (None), and weighted C2C datasets. Weights were estimated using a logistic (Log.), covariate balancing propensity score (CBPS), entropy balancing (EB), or random forest (RF) method. Continuous covariates were summarized by mean (standard deviation) and categorical covariates by proportion. Standardized difference in means relative to NHANES-REP are presented in the figure on the right [108]. Propensity weights were estimated with missing values in the C2C imputed once. Covariates are described in Section 3.4.1.

and 95% confidence intervals using the analytic variance estimates for each racial/ethnic group for each of the 9 responses. We report the full results with no weighting and each of the four types of propensity weights in Appendix B (Table B.1). A selection of these results are depicted using forest plots to discuss the impact of weighting.

Across all forest plots (Figure 3.4), we observed that the standard errors increased with weighting, but this added variance better reflects the true uncertainty in the estimates and their ability to generalize to an external population. For example, the odds ratio comparing Hispanics to NH Whites for willingness to be contacted about studies with lumbar puncture (LP) had a noticeably wider confidence interval for the weighted estimates. The C2C underrepresents Hispanic subjects relative to the US population and the wider confidence intervals reflect this lack of information on the subpopulation. Several statistically significant odds ratios were no longer significant after incorporating estimated propensity weights. NH Asians had significantly higher odds of being willing to be contacted about studies involving LP compared to NH Whites in the original analysis, but this relationship was no longer statistically significant after weighting. Salazar et al. (2020) [100] found it surprising that NH Asians would be more willing to undergo an LP because previous studies had found them less willing relative to NH Whites [84]. They speculated that NH Asians in the C2C had been exposed to more education about the LP procedure through outreach events for older Chinese adults. Accounting for sampling bias with estimated weights has attenuated this relationship to the null which aligns with previous findings.

The models using logistic and CBPS estimated weights tended to have similar estimates and confidence intervals. The CBPS model uses a logistic regression model but incorporates moment balancing conditions into the model fitting. These additional constraints did not impact the final result substantially when compared to the standard logistic regression derived estimates. The estimates using RF and EB weights had high variability and differed from the results using logistic and CBPS weights. Additionally, the point estimates from

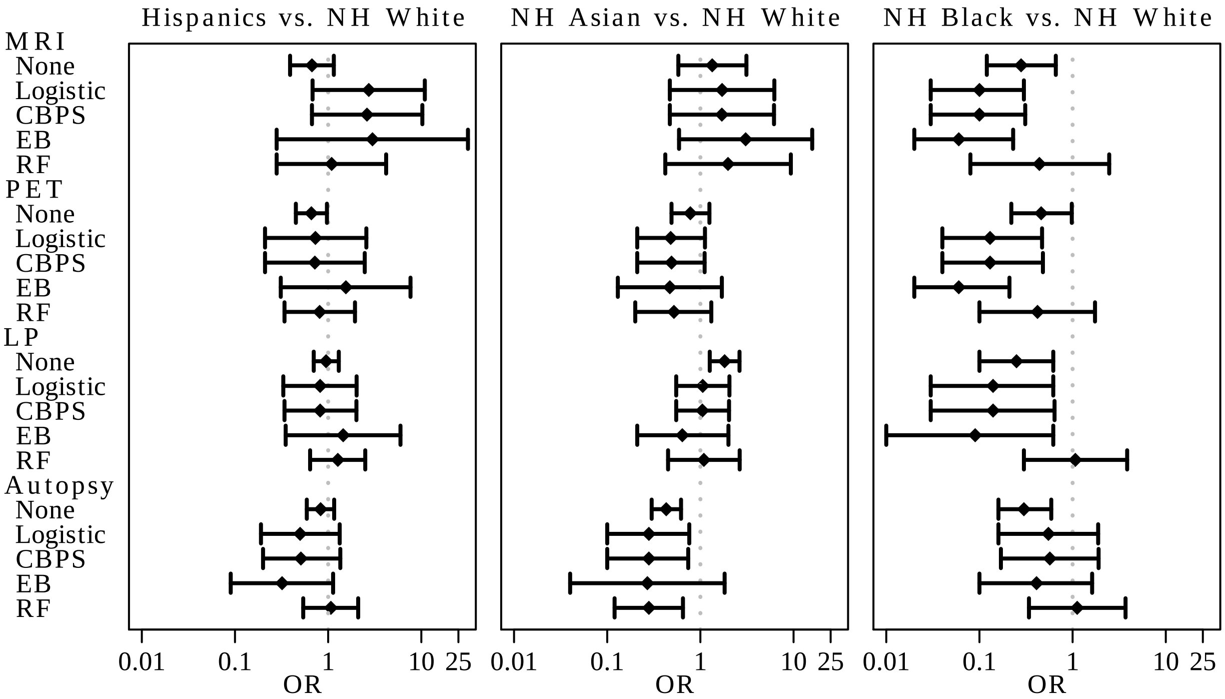


Figure 3.4: Forest plots of the estimated odds ratios (OR) and 95% confidence intervals for the racial ethnic differences analysis with MRI, PET, lumbar puncture (LP) and autopsy as the response. Results are presented for unweighted analysis (None) along with the weighted analysis using propensity weights estimated with logistic, covariate balancing propensity score (CBPS), entropy balancing (EB), and random forest (RF) methods.

the models using RF weights tended to differ the most from the other 3 weighted models. Random forest is unique because it is both non-parametric and does not target covariate balancing. An advantage of decision trees is they naturally include interactions in modeling, but in sparse data with little representation of subpopulations this can lead to increased variability. Weight trimming, where estimated probabilities of 0 or 1 were replaced with 0.01 and 0.99, may have also impacted bias and variance estimates as well as the population of inference [67].

Incorporating estimated propensity weights not only impacted the uncertainty, it also changed the direction of point estimates. For example, in the original analysis Hispanics had lower odds of being willing to be contacted about studies that involve an MRI scan, but most of the weighted point estimates suggested Hispanics may actually have higher odds. The results for the original analysis were close to being statistically significant but the weighted models showed little evidence of an effect which leads to a different interpretation of the results. The four weighted point estimates were not covered by the unweighted confidence interval.

3.5 Discussion

Our results demonstrated the utility of using convenience samples in concert with a representative sample to estimate weights that can be used to estimate population representative parameters of interest. Convenience samples are widely available and often used in research studies, but failing to account for the selection mechanism can lead to biased estimates and underestimation of the true variance of estimates. It is important to carefully select a target population and design studies and analyses that generalize to this population. If researchers are not able to obtain a representative sample because of ethical or practical considerations, they are forced to use a convenience sample. Since estimated propensity weights can only balance a convenience sample on observed covariates, researchers must take care to collect any covariates that they hypothesize are related to both the outcome and sampling probability. Additionally, any subpopulation with a convenience sample membership probability of zero and thus not represented in the convenience sample cannot be included in the target population. Researchers must carefully consider which covariates should be collected and which subpopulations are being sampled into a convenience sample to allow for valid estimates of associations in the desired target population.

In the analysis of racial/ethnic differences of research willingness, weighted confidence intervals were generally at least twice as wide as unweighted confidence intervals. Incorporating propensity weights can increase the variability of parameter estimates because subjects with a low estimated probability of convenience sample membership will have large estimated weights and undue influence on the estimated associations [71]. Although propensity weights may increase variance, they can reduce bias of the estimated association in the population of interest. Using an unrepresentative sample provides less information about the target population, and thus the increased variance of our estimator reflects this uncertainty [75]. To measure how much the sampling mechanism impacts efficiency, we can compute the design effect, which is the ratio of the variance of the parameter estimate in an unrepresentative sample compared to a simple random sample [75, 64]. When comparing empirical variances from the simulation study for logistic-weighted coefficient estimates fit in a biased sample compared to unweighted estimates in a SRS, the design effect ranges from 1 to 1.6. Thus, we will need a biased sample that is up to 1.6 times bigger than a simple random sample to obtain the same variance.

For estimating the variance of parameter estimates in the outcome model, we compared a resampling approach and our proposed analytic approach for a logistic regression model that accounts for uncertainty arising from the propensity weight estimation. In empirical studies, we found the proposed analytic estimates performed better than the bootstrap estimate even though the analytic estimate does not account for the model selection when estimating propensity weights. Surprisingly, the bootstrap estimate tends to overestimate the uncertainty even though we replicated the estimation method within each bootstrap sample. Perhaps there was more variability in estimated weights within bootstrap samples than within the full simulated data set. Previous work on implementing entropy balancing weights [110] and exact matching using the propensity score [10] also reported conservative bootstrap variance estimates, but the stratified double bootstrap where units are resampled from the survey sample and convenience sample used by [2] provides similar variance estimates to a design-based approach. The proposed analytic variance for model-based propensity weight estimation methods accounts for uncertainty in the estimated weights. The design based standard errors, however, perform well even though they fail to account for the propensity weight estimation process. We suggest using the proposed variance estimator with a model based propensity weight estimation procedure because it may perform better than the design based estimate, but the design based approach should perform well if needed. In our context, the two methods did not diverge substantially, but they could if there is high variability in the propensity weights. It is possible to derive an analytic variance estimator that accounts for the weight estimation for the CBPS model. However, it is made difficult because the CBPS model is overspecified and fit using generalized method of moments [61, 53]. One might be able to incorporate the final scientific model into the CBPS model as an additional balancing constraint for a simultaneous estimation approach. This may be an interesting area of research to pursue.

All four propensity weight estimation models decreased bias in the simulation study. Algorithmic propensity weight estimation methods are very flexible, but random forest provided the smallest degree of bias reduction and the largest variance in the simulation study. Using RF-derived weights provided the poorest covariate balance (Figure 3.3). Models that incorporate covariate balancing into model-selection help ensure covariate balance in the biased sample. EB, however, scales better than forward step-wise model selection for logistic regression, but CBPS tends to be slower due to the additional constraints. Estimates using CBPS did not deviate substantially from those using logistic regression, so the additional balancing constraints did not improve performance. We used the default settings for the `CBPS` package and users can change the settings to focus more on covariate balance. Although EB balanced covariates better than the logistic model in the applied example, they both reduced bias in estimated associations to the same extent in simulation study. Likelihood-based regression models, such as logistic regression, allow for an analytic variance estimate that fully accounts for uncertainty. It can also be easily expanded to include interaction and smoothers to allow for greater flexibility, but the second order terms seemed to perform well enough in our experiments. In practice, we suggest using logistic regression because it effectively reduces bias under our assumptions, is familiar to many scientists, is broadly accessible in different statistical software packages, and allows for an analytic variance estimate that accounts for uncertainty from estimation of propensity weights. The prediction models we implemented are only several examples of many different options one could use. Practitioners can select their preferred prediction model.

NHANES is a practical choice for generalizing biomedical studies to the US population. Different research areas may, however, collect variables that are not recorded in NHANES but are believed to be strongly related to the sampling probability. Other national surveys collect different variables and may be more relevant to different research areas. Other examples of national surveys are the American Community Housing Survey which collects population and housing information, the Behavioral Risk Factor Surveillance System which conducts health-related telephone interviews, the General Social Survey which studies American society, and the Current Population Survey that collects labor force statistics. Additionally, researchers may want to generalize to a population outside of the US. If the target population is a subset of the US population, NHANES can be subset and used as the representative sample. Otherwise, other representative samples need to be obtained. Researchers can consider census data if available, government sponsored national surveys, or international surveys. After specifying the target population, one should consider which samples are most representative and accessible.

Overall, estimated propensity weights reduce bias on parameter estimates from unrepresentative sampling. We, of course, are unable to account for any unmeasured covariates that may contribute to selection bias. Additionally, we are unable to learn about subpopulations that were never sampled. For example, if there are no NH Blacks with less than a high school education in the C2C, we cannot weight this missing subpopulation. We collapsed different variables to match across different datasets and were unable to empirically evaluate if these are equivalent definitions. Implementing estimated propensity weights increases the uncertainty of estimates but this reflects the information available on target population parameters.

Convenience samples are easily collected and are used for research in many disciplines. The NHANES dataset is a rich, open access dataset that will likely have many overlapping covariates with convenience samples. Estimated propensity weights using NHANES is practical and effective at addressing selection bias concerns in convenience samples when trying to generalize to the non-institutionalized US population.

Chapter 4

Propensity Scores in Convenience Samples

4.1 Introduction

In the previous chapter, we discusses a method for estimating sampling weights. In this chapter, we apply estimated sampling weights to estimate causal effects for observational studies with a biased sample. Please note that in the previous chapter we used the term "propensity weights" to refer to inverse probability of sampling weights, but in the remainder of the dissertation we use the term "sampling weights" to avoid confusion with the propensity for treatment.

Convenience samples are often used to estimate causal effects in observational studies, but the resulting effect estimates are subject to potential confounding and sampling bias. One example of a convenience sample is the National Alzheimer's Coordinating Center (NACC) Uniform Data Set. NACC collects longitudinal demographic, clinical and specimen data on participants from 41 Alzheimer's Disease Research Centers[13]. Although there has been an effort at some centers to recruit more representative cohorts, most of the volunteers in the NACC data set come from referrals and so it is not a random sample of older adults in the United States (US) population. Highly educated and non-Hispanic (NH) white volunteers are overrepresented in the NACC sample as is common in clinical research [87]. Suppose we want to use the NACC data set to estimate the effect of vitamin E supplementation on the Functional Activities Questionnaire (FAQ) [81], a measure of activities of daily living, among cognitively normal older adults in the US. Then we must account for the NACC selection mechanism as well as any potential confounders to avoid biased estimates of the effect of vitamin E supplementation. We will accomplish this by studying methodology for using convenience samples to estimate causal effects for a prespecified target population.

Using propensity scores to address potential confounding is a common method for estimating causal effects in observational studies. Adjusting for propensity scores alone, however, does not address sampling bias. Inverse probability of sampling weights are commonly used to obtain generalizable inference for survey samples where the sampling probabilities are prespecified by design [75]. Sampling weights for convenience samples can be estimated by using a more representative sample [15, 26, 2, 94, 40, 117, 88]. Incorporating sampling weights into a propensity score analysis is not straight forward because both the propensity score estimation model and the causal model can be weighted when they are estimated, but there is debate in the literature about whether the propensity score model should be weighted [93, 119, 32, 9, 68]. The objective of this chapter is to quantify when and if propensity score and outcome models should be weighted with sampling weights when using convenience samples and to quantify uncertainty in treatment effect estimates.

Ridgeway et al. (2015) studied the impact of including sampling weights in the propensity score and outcome models for survey samples with known sampling weights [93]. They conclude that the outcome model should always be weighted and in addition the propensity score model should be weighted under 3 scenarios: (1) When there is a covariate used in the sampling weight calibration that is not available in the data set for estimating propensity scores, (2) when residual confounding occurs from lack of flexibility in the propensity score model, and (3) when data from multiple survey samples is combined and thus sampling weights are based on different covariates. More recent papers have disagreed with Ridgeway et al. and suggested that only the outcome model should be weighted, but they assume the propensity score model is correctly specified [9, 68].

In this chapter, we will focus on the second scenario and quantify how well we are able to estimate a marginal treatment effect estimate for a target population when there are heterogenous treatment effects in under- or over-represented subpopulations in a convenience sample. Ridgeway et al.'s work assumes that sampling weights are pre-specified or contain all known information related to the selection mechanism. In this chapter, we will investigate the impact and feasibility of estimating sampling weights for convenience samples to account for sampling bias in estimated treatment effects. We use Monte Carlo methods to quantify the bias reduction when sampling weights are calibrated using covariates related to sampling bias and compare it to when the covariates are not available, but proxies associated with them are instead available. Lastly, we consider the impact of estimating sampling weights and propensity scores on the uncertainty in the final treatment effect estimate and propose a variance estimate that accounts for the estimation of both quantities.

The remainder of this chapter is organized as follows. In Section 4.2 we formalize the method for estimating sampling weights, propensity scores, and the treatment effect to account for the selection mechanism. We also derive the variance estimate that accounts for uncertainty in when simultaneously estimating sampling weights and propensity scores. In Section 4.3 we present Monte Carlo studies to compare the impact on bias when including sampling weights in the propensity score model, outcome model, or both when the selection mechanism is known. We then quantify the impact on bias when those variables that determine sampling probabilities are unobserved but where proxy variables of varying strength are observed.

Next, we assess the performance of the proposed variance estimate relative to the empirical error. In Section 4.4 we apply the methods from this chapter and estimate sampling weights for participants in the NACC using a representative US population sample obtained from the National Health and Nutrition Examination Survey (NHANES). We use inverse probability of sampling weighted propensity scores to estimate the effect of Vitamin E on the ability to perform daily activities for cognitively normal older adults. We compare the estimated effect when omitting sampling weights and including them in the propensity score model, outcome model, or both. We conclude with a discussion about the advantages and disadvantage of weighting both models with estimated sampling weights in Section 4.5.

4.2 Methods

In this paper, we implement an analogous approach for estimating sampling weights for convenience samples via a representative data set as discussed in our previous work. [15] We then implement the estimated sampling weights to estimate a marginal treatment effect in a target population. To estimate a causal effect with a convenience sample, we need to collect two samples: a convenience sample and a representative sample for estimating sampling weights. Both of these samples have specific variables they must contain. First, the convenience sample must include the treatment indicator, the response, covariates related to potential confounding, and covariates related to the sampling probability. Second, the representative sample only needs to contain covariates related to the sampling probability.

To construct a data set for analysis, we need two collections of variables, $\mathcal{X}_R$ and $\mathcal{X}_C$, which represent that variables collected in the representative and convenience samples, respectively. Next, we will formalize which variables need to be included in $\mathcal{X}_R$ and $\mathcal{X}_C$. Consider the set of variables collected in both of these sets, so that $\mathcal{V} = \mathcal{X}_R \bigcup \mathcal{X}_C$. $\mathcal{V}$ should contain variables related to the sampling probability and will be used to estimate sampling weights. The other variables necessary to estimate a causal effect, the treatment $\mathcal{A}$, outcome $\mathcal{Y}$ and potential confounders $\mathcal{Z}$ are all collected as part of the convenience sample. So $\mathcal{A} \subset \mathcal{X}_C$, $\mathcal{Y} \subset \mathcal{X}_C$, and $\mathcal{Z} \subset \mathcal{X}_C$. There may be some overlap between the covariates needed to estimate a causal effect ($\mathcal{A}$, $\mathcal{Y}$, and $\mathcal{Z}$) and the variables collected in the representative sample $\mathcal{X}_R$, but we assume that the majority of these variables are only collected in the convenience sample or $(\mathcal{A} \bigcup \mathcal{Y} \bigcup \mathcal{Z}) \setminus \mathcal{X}_R \neq \emptyset$. Since we are unable to directly estimate a causal effect in the representative sample, our goal is to leverage auxiliary information from the representative sample to estimate sampling weights that will let us estimate a causal effect for a target population using a convenience sample. To do this, we will estimate propensity scores in a convenience sample using a model weighted by estimated sampling weights, and then estimate a causal effect with estimated propensity score and estimated sampling weights.

4.2.1 Assumptions

To estimate causal effects using a convenience sample we must make two sets of assumptions. The first set involves assumptions necessary for estimating a causal effect and the second set includes assumptions for valid estimation of sampling weights.

As discussed in Chapter 2, there are three assumptions necessary for making causal conclusions using propensity scores: unconfoundedness, positivity, and the stable unit treatment value assumption (SUTVA). These assumptions are discussed in Chapter 1 and 12 of Imbens and Rubin (2015) [62] and Appendix 1 of Greenland, Robins, and Pearl (1999)[50]; among other places. The unconfoundedness assumption states that the potential outcomes are independent of the treatment assignment conditional on the observed covariates. Next, the positivity assumption requires each unit to have positive probability of being in both treatment groups. Let A be an indicator for being in the treatment group and Z denote observed covariates, then the positivity assumption can be expressed $0 < \Pr(A = 1|Z = z) < 1$ for all units. And lastly, SUTVA states that a unit's treatment assignment does not impact the potential outcomes of any other unit and there is no hidden variability in the treatment. These three assumptions are required to estimate causal effects using propensity scores.

As discussed in our previous chapter on estimating sampling weights[15], similar assumptions are needed to use estimated sampling weights to address sampling bias. We need to assume a version of unconfoundedness, that the response is independence of the selection probability, conditional on observed covariates. This means that all covariates related to both the sampling probability and response must be collected in both the representative and convenience samples. More formally, $\mathcal{V}$ must contain all variables associated with the sampling probability and response. We must also assume a form of positivity, all subjects must have a positive probability of being sampled into the convenience sample. Let V be the observed covariates specified in $\mathcal{V}$ and C be an indicator of being in the convenience sample, $0 < \Pr(C = 1|V = v) < 1$ for all subjects in the representative and convenience samples.

4.2.2 Estimating Causal Effects in a Convenience Sample

Let A_i be the indicator of receiving treatment ($A_i \in \{0, 1\}$), then for individual i who is exposed to treatment $A_i = 1$, their potential outcome is $Y_i(1)$. If subject i is exposed to treatment $A_i = 0$, then their potential outcome is $Y_i(0)$ [95]. Since we are assuming two treatment levels, each subject has two potential outcomes at a given time corresponding to the two treatment levels. A causal effect is defined as a comparison in the potential outcomes for two treatment groups for subject i. We are only able to observe one potential outcome per individual at a given time, because they can only be in one treatment group. The observed response for individual i exposed to treatment $A_i = a_i$ is then

$$Y_i = (1 - a_i)Y_i(0) + a_iY_i(1)$$

Ideally, we would calculate an individual's treatment effect, such as $Y_i(1) - Y_i(0)$, but we are unable to do so because only one of these terms is observable at a given time point. Instead to identify a causal effect, we can estimate the average causal effect by comparing the mean outcomes in the treated and control groups. In this paper we will focus on the average treatment effect (ATE)

$$E_P[Y(1) - Y(0)] \tag{4.1}$$

where E_P denotes an expectation over some population, P. In an observational study, individuals' characteristics in the control group may be systematically different from those in the treated group because researchers do not control the treatment assignment mechanism. There may be confounders related to the treatment assignment and the potential outcomes with different distributions in the treatment and control groups. To address this, we estimate the ATE when comparing individuals with a similar propensity for treatment by conditioning on the propensity score. Let z denote a vector of observed covariates that should include variables related to the treatment probability. The propensity score $e(z)$ is then the probability of being exposed to treatment $A = 1$ conditioned on Z, or $e(z) = \Pr(A = 1|Z = z)$. To account for potential confounding we can estimate the ATE by conditioning on propensity scores, or

$$E_P\big[E_Z\big\{Y(1) - Y(0)|e(z_i)\big\}\big]. \tag{4.2}$$

This expectation is over a specific population P. In this chapter, we address the scenario where the distribution of the Z and Y in the convenience samples differs from the distribution in the population P because it is not a random sample. Our target estimand is the average treatment effect in a target population defined in Equation 4.2 and not the average treatment effect conditioned on being sampled. Let C be an indicator of being sampled into the convenience sample. If we ignore the selection mechanism instead of estimating the ATE in the population (Equation 4.1), we will instead estimate the ATE in the sampled population,

$$E_{P|C}[Y(1) - Y(0)|C = 1]. \tag{4.3}$$

In the absence of confounding, we could account for the selection mechanism to estimate the population ATE by estimating sampling weights, w, such that

$$E_{P|C}[w(Y(1) - Y(0))|C = 1] = E_P[Y(1) - Y(0)]. \tag{4.4}$$

We define the form of these weights and discuss estimation strategies in the following section. In this chapter we discuss combining the approach for using propensity scores to address confounding in Equation 4.2 and the approach for using sampling weights to address the selection mechanism in Equation 4.4. Weighting the propensity score estimation model and weighting the treatment effect with estimated sampling weights for a convenience sample will allow us estimate our estimand of interest from Equation 4.2.

When estimating a causal effect using propensity scores in a representative sample, there are two steps. First, you (1) estimate propensity scores and then you (2) estimate an effect conditioning on the propensity score. There are several approaches to condition on the propensity score including matching, stratification, adjustment, and weighting (for an overview see Chapter 12 of Imbens and Rubin 2015) [62]. In this chapter, we will focus on propensity score adjustment in regression models as an example because it allows us to derive an analytic variance estimate, but all of these options estimate the average treatment effect [8]. Our findings about whether the propensity score estimation model should be weighted will likely extend to the other propensity score methods.

Estimating causal effects in a convenience sample without accounting for the selection mechanism can lead to biased estimates of the effect in a target population. Incorporating estimated sampling weights into the analysis allows us to estimate causal effects for a target population. We will then have three steps, (1) estimate sampling weights, then (2) estimate propensity scores, and (3) estimate a causal effect. Estimated sampling weights can be incorporated into the analysis by weighting the propensity score estimation model, weighting the outcome model, or weighting both models. The goal of this chapter is to compare these three options for incorporating sampling weights and identify when weighting is necessary. We will start by laying out the analysis plan when both the propensity score and outcomes models are weighted, and then illustrate the differences when they are omitted.

Sampling weight model

Let $\mathcal{C}$ and $\mathcal{R}$ be sets of individuals in a convenience sample and representative sample, respectively. When estimating sampling weights, we subset the representative and convenience samples down to the variables collected in both data sets, $\mathcal{V}$ and concatenate them into a combined data set. We construct an indicator of convenience sample membership, C_i, which equals one for subjects from the convenience sample and zero for those from the representative sample. Let v_i be a $1 \times M$ vector of covariates (including 1 to allow for estimating an intercept) for the sampling weight estimation model and γ be a $M \times 1$ vector of parameters for the sampling weight estimation model. To estimate the sampling probability, we use logistic regression to estimate $p_i = \Pr(C_i = 1|v_i)$. There are many options of predictive models that can be used to estimate the sampling probability, but we used logistic regression with forward stepwise model selection with Akaike information criterion (AIC) for computational tractability. We found this approach generally performed well in Chapter 3 [15].

We use a generalized linear model (GLM) framework to fit the logistic regression model. [85] Define $\text{logit}(\cdot)$ as the logit link function and Ψ_i as a linear predictor, the sampling weight estimation model is then given by

$$\Psi_i = \text{logit}(p_i) = v_i\gamma$$

with the corresponding score or estimating equation for the m-th parameter,

$$T_m(\gamma) = \sum_{i \in \mathcal{C} \cup \mathcal{R}} T_{mi}(\gamma) = \sum_{i \in \mathcal{C} \cup \mathcal{R}} (C_i - p_i) v_{im} = 0. \tag{4.5}$$

Note that p_i is a function of γ.The estimated sampling probabilities, $\hat{p}_i$, are then used to formulate the estimated sampling weights,

$$\hat{w}_i \propto \frac{1 - \hat{p}_i}{\hat{p}_i}$$

where the weights are scaled so that $\sum_{i \in \mathcal{C}} w_i = 1$.

Propensity score model

Let A_i be an indicator of whether subject i received treatment and $e(z_i) = \Pr(A_i = 1 | Z_i = z_i)$ be the propensity score for subject i. We define z_i as a $1 \times L$ vector of covariates and ξ as a $L \times 1$ vector of parameters for the propensity score estimation model including an intercept. Although there are many options for which model to estimate a propensity score because it is commonly used in practice, can be made very flexible with the addition of higher order terms, interactions or smoothers, and is easy to implement. The logistic model for estimating propensity scores with linear predictor Φ_i is,

$$\Phi_i = \text{logit}(e(z_i)) = z_i \xi.$$

The unweighted estimating equation for the l-th parameter, where $e(z_i)$ is a function of ξ, is

$$S_l(\xi) = \sum_{i \in \mathcal{C}} S_{li}(\xi) = \sum_{i \in \mathcal{C}} (a_i - e(z_i)) z_{il} = 0 \tag{4.6}$$

and the weighted estimating equation is

$$\bar{S}_l(\xi) = \sum_{i \in \mathcal{C}} \bar{S}_{li}(\xi; \hat{w}) = \sum_{i \in \mathcal{C}} \hat{w}_i(a_i - e(z_i))z_{il} = 0. \tag{4.7}$$

Implementing sampling weights into the analysis by using $\bar{S}_l(\xi)$ in Equation 4.7 will calculate estimated propensity scores, $\hat{e}(z_i)$. Ignoring the selection mechanism and fitting the unweighted estimating equation S_l in Equation 4.6 will calculate estimated propensity scores that will correspond to the probability of receiving treatment in the sampled population. We will refer to the unweighted propensity score estimates as $\hat{e}^*(z_i)$.

Outcome model

Finally, we are able to estimate a causal effect. We use a generalized linear model framework to estimate a causal effect to allow for different types of responses [85]. We will start by considering the models that use the estimated propensity score $\hat{e}(z_i)$. Let Y_i be the response variable, $\mu_i = E[Y_i]$ be the expectation of the response, and $V(\mu_i) = \text{Var}(Y_i)$ be the variance of the response. Additionally, $g(\cdot)$ is the link function and η_i is a linear predictor. Then $x_i = [1 \quad a_i \quad \hat{e}(z_i)]$ are the covariates for subject i in the outcome model and β is the 3×1 vector of parameters. The outcome model using estimated weighted propensity scores is

$$\eta_i = g(\mu_i) = \beta_0 + \beta_1 a_i + \beta_2 \hat{e}(z_i)$$

The unweighted estimating equation is

$$U_j(\beta) = \sum_{i \in \mathcal{C}} U_{ji}(\beta) = \sum_{i \in \mathcal{C}} \frac{(y_i - \mu_i)}{V(\mu_i)} \left[\frac{\partial \eta_i}{\partial \mu_i}\right]^{-1} x_{ij}$$

which provides the estimate of the treatment effect when only the propensity score model is weighted, $\widetilde{\beta}_1$. The weighted estimating equation is

$$\bar{U}_j(\beta) = \sum_{i \in \mathcal{C}} \hat{w}_i U_{ji}(\beta) = 0 \tag{4.8}$$

which results in the doubly-weighted estimate of the average treatment effect, $\hat{\beta}_1$. The analogous model when using the unweighted propensity score is

$$\eta_i^* = g(\mu_i^*) = \beta_0^* + \beta_1^* a_i + \beta_2^* \hat{e}^*(z_i)$$

The unweighted estimating equation using the unweighted propensity score and $x_i^* = [1 \quad a_i \quad \hat{e}^*(z_i)]$ is

$$U_j(\beta^*) = \sum_{i \in \mathcal{C}} U_{ji}(\beta^*) = \sum_{i \in \mathcal{C}} \frac{(y_i - \mu_i^*)}{V(\mu_i^*)} \left[\frac{\partial \eta_i^*}{\partial \mu_i^*}\right]^{-1} x_{ij}^*$$

which provides the naive estimate of the treatment effect that excludes sampling weights and thus ignores the selection mechanism, $\widetilde{\beta}_1^*$. The weighted estimating equation that uses unweighted propensity score estimates is

$$\bar{U}_j(\beta^*) = \sum_{i \in \mathcal{C}} \hat{w}_i U_{ji}(\beta^*) = 0$$

which results in the estimate of the average treatment effect that only accounts for sampling bias in the outcome model, $\hat{\beta}_1^*$.

4.2.3 Analytic Variance Estimator

The standard analytic sandwich estimator of the variance of parameter estimates in generalized linear models that accounts for survey weights is an extension of the Huber-White variance estimate [41]. This variance estimator is provided in the `svyglm` function in the `survey` package[75, 77] in `R`. The standard analytic variance estimator assumes that sampling weights $\hat{w}$ and propensity scores $\hat{e}(z)$ used in the outcome model are fixed. To derive an analytic variance estimator that accounts for the uncertainty from estimating sampling weights and propensity scores, we propose to use a simultaneous estimating equation approach similar to the one used by Schildcrout and Rathouz (2010) [103]. The goal is to treat the three estimating equations: the sampling weight estimating equation $T_m(\gamma)$ (Equation 4.5), the propensity score estimating equation $\bar{S}_l(\xi)$ (Equation 4.7), and the outcome model estimating equation $\bar{U}_j(\beta)$ (Equation 4.8). In this section, we update our notation to emphasize that the propensity score model is a function of the estimated sampling weights and the outcome model is a function of the estimated propensity scores and sampling weights and call the estimating equations $\bar{S}_l(\xi, \gamma)$ and $\bar{U}_j(\beta, \xi, \gamma)$, respectively. We treat these estimating equations as if they are jointly estimated giving us a stacked estimating equation:

$$\kappa = \begin{pmatrix} \sum_{i \in \mathcal{C} \cup \mathcal{R}} T_i(\gamma) \\ \sum_{i \in \mathcal{C}} \bar{S}_i(\xi, \gamma) \\ \sum_{i \in \mathcal{C}} \bar{U}_i(\beta, \xi, \gamma) \end{pmatrix} = 0$$

We can use a Taylor series expansion of the stacked estimating equation (see Appendix C) to obtain the following variance estimator of the parameter estimates

$$\hat{V}_{Prop}[(\hat{\gamma}, \hat{\xi}, \hat{\beta})] = \hat{I}^{-1} \hat{Q} \hat{I}^{-1}, \tag{4.9}$$

where the hats denote estimates. Q is the true variance of the stacked estimating functions κ,

$$Q = \text{Var}(\kappa) = E(\kappa \kappa^T).$$

A method of moments estimator for Q is $\hat{Q} = \hat{\kappa}\hat{\kappa}^T$. I is Fisher's information matrix

$$I = \begin{pmatrix} I_{TT} & 0 & 0 \\ I_{ST} & I_{SS} & 0 \\ I_{UT} & I_{US} & I_{UU} \end{pmatrix}. \tag{4.10}$$

Definitions of the terms in I and corresponding estimates are detailed in Appendix C.

4.2.4 Implementation of Variance Estimator

We developed an `R`[91] package to estimate causal effects when are both the propensity score model and outcome model are weighted with estimated survey weights. It also returns standard error estimates using our proposed variance estimator described in this section. Description of the `estweight` package and with instructions for downloading and use can be found at (`https://github.com/oliviabern/estweight`). In short, the `convPS` function requires a convenience sample and representative sample. The user specifies which variables are used for estimating sampling weights and propensity scores, along with the names of the treatment and response variable. The function then returns the estimated causal effect and corresponding standard error estimate.

4.3 Simulation Study

We conducted a simulation study to investigate the impact of including sampling weights in the propensity score and outcome models. Our goal is to estimate the causal effect for the target population using a convenience sample. We simulated data for an unrepresentative sample and compared estimated causal effects fit in an unrepresentative, or convenience, sample and compared them to estimates from a simple random sample (SRS) from the target population. We estimate bias of the estimated treatment effect by the difference between the mean estimate using the convenience sample and the mean estimate from the SRS across simulations. Code for this simulation study is available at `https://github.com/oliviabern/weighted_propscores`.

4.3.1 Simulation Scenario

Suppose there is an indicator, K_i, for a subpopulation that is overrepresented in the convenience sample relative to the target population, but is unmeasured. Instead we observe a covariate X_{1i} that is correlated with K_i. We let K_i and X_{1i} be binary variables that are transformations of normally distributed covariates K_i^* and X_{1i}^* with correlation ρ. To reflect this, let

$$\begin{pmatrix} K_i^* \\ X_{1i}^* \end{pmatrix} \sim \mathrm{N}\Bigg(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \Bigg).$$

We transform K_i^* and X_{1i}^* with a probit link so

$$K_i = \begin{cases} 1, & \text{if } K_i^* > 0 \\ 0, & \text{if } K_i^* \leq 0 \end{cases}$$

$$X_{1i} = \begin{cases} 1, & \text{if } X_{1i}^* > 0 \\ 0, & \text{if } X_{1i}^* \leq 0 \end{cases}.$$

Individuals are sampled with uniform probability regardless of K into the simple random sample, but the probability of being sampled into the convenience sample is

$$\Pr(\text{Sampled}_i) = .6K_i + .2 \tag{4.11}$$

so individuals with $K = 1$ are four times more likely to be sampled than individuals with $K = 0$. Let $X_{2i} \sim \mathrm{N}(0, 2)$ and $X_{3i} \sim \mathrm{N}(0, 2)$. Sampling bias can lead to bias in the estimated propensity scores if the probability of receiving treatment is dependent on the probability of being sampled. Let $Z_i = [X_1, X_2, X_3, X_4]$ and suppose the covariates X_2 and X_3 have a differential relationship with the probability of receiving treatment $e(Z_i)$ for people in the $K = 0$ and $K = 1$ subpopulations, so that

$$e(Z_i) = \mathrm{expit}\Big\{(1 - K_i)(log(\alpha_{02})X_{2i} + log(\alpha_{03})X_{3i}) + K_i(log(\alpha_{12})X_{2i} + log(\alpha_{13})X_{3i})\Big\}. \tag{4.12}$$

Let A_i be the treatment indicator, so that $A_i \sim \mathrm{Bernoulli}(e(Z_i))$. Unrepresentative sampling can lead to bias in the estimated treatment effect if there is a differential treatment effect in subpopulations that are disproportionately sampled. To illustrate this, we simulate a differential treatment effect for the $K = 1$ and $K = 0$ subpopulations. Let Y_i be the response, $X_{4i} \sim \mathrm{N}(0, 1)$, and $X_{5i} \sim \mathrm{N}(0, 1)$. The response $Y_i \sim \mathrm{N}(\mu_i, 1)$ and

$$\mu_i = \delta_A A_i + \delta_{AK} A_i \times K_i + \delta_2 X_{2i} + \delta_3 X_{3i}^2 + \delta_4 X_{4i} + \delta_5 X_{5i}^3 \tag{4.13}$$

Within each simulated data set we generated a finite population with 10,000 observations and then from this we draw a convenience sample of size 2000 according to the sampling probabilities in Equation 4.11 and a representative sample of size 2000 where each individual has an equal probability of selection. Let C_i be an indicator of being in the convenience sample and $V_i = [X_{1i}, ..., X_{5i}]$ be the covariates used for estimating the sampling weights. We estimated the probability of being sampled $\Pr(C_i = 1 | V_i = v_i)$ relative to the representative sample for the convenience sample using logistic regression with forward-selection with AIC. We included V with first and second order terms in the model scope, but excluded interactions between covariates from the scope. When the correlation, ρ, between K and X_1 is 1, then $X_1 = K$ which reflects the scenario where the covariate K related to the sampling probability is observed which is analagous to the situation considered by Ridgeway et al. [93]. The estimated sampling weights are then

$$\hat{w}_i = \frac{1 - \widehat{\Pr}(C_i = 1|V_i = v_i)}{\widehat{\Pr}(C_i = 1|V_i = v_i)}.$$

Let $Z_i = [X_{1i}, ..., X_{5i}]$ be the covariates used for estimating the propensity score. Within the convenience sample, we estimated propensity scores $\hat{e}(z_i) = \Pr(A_i = 1|Z_i = z_i)$ using logistic regression with forward-selection using AIC with Z_i and included linear terms, quadratic terms, and interactions in the model scope. We used the design based AIC (dAIC) [76] when including sampling weights in the propensity score model. We then estimated the estimated treatment effect by adjusting for the propensity score, where

$$g(\mu_i) = \beta_0 + \beta_1 a_i + \beta_2 \hat{e}(z_i).$$

In our case where Y_i is normally distributed, we used the canonical link so $g(\mu_i) = \mu_i$. In the convenience sample, we estimated propensity scores and a treatment effect in the same way. We compared estimates of β_1 from the convenience sample when including or omitting sampling weights in both the propensity score and outcome models to estimates from the representative sample. Notice that our estimand of interest is β_1, the marginal treatment effect even though there is treatment effect heterogeneity in Equation 4.13.

We conducted 1000 simulations and compared the bias in estimates fit in the convenience sample relative to estimates from the representative sample. Unless otherwise specified, parameters in Equation 4.12 are $\alpha_{02} = 1.3$, $\alpha_{12} = 2$, $\alpha_{03} = 0.4$, and $\alpha_{13} = 1.5$ and parameters in Equation 4.13 are $\delta_A = 1$, $\delta_{AK} = 3$, $\delta_2 = 1.5$, $\delta_3 = -2$, $\delta_4 = -1$, and $\delta_5 = 1.5$. We started with the scenario where K is observed or $\rho = 1$. We compared the mean absolute bias relative to the representative sample when varying the true coefficients in the outcome model (δ_A, δ_{AK}, δ_2, δ_3, δ_4 and δ_5). We also evaluated the mean absolute bias relative to the SRS when varying the parameters in the true propensity model (α_{02}, α_{03}, α_{12}, and α_{13}). Next, we examined the relative bias when the correlation ρ decreased to show the impact of failing to fully measure the variable related to the sampling probability. Lastly, we assessed the uncertainty of the parameter estimates. We fixed the coefficients in the outcome model and compare the empirical standard errors across the weighted and unweighted models. Finally, we compare the standard error estimates (naive non-robust, standard robust, and the proposed robust estimates) to the empirical Monte Carlo estimate.

4.3.2 Simulation Results

In Figure 4.1, we demonstrated the impact of each term in Equation 4.13 which is the data generating mechanism for the response by varying the coefficients δ_A, δ_{AK}, δ_2, δ_3, δ_4, and δ_5. We report the absolute bias in the estimated treatment effect relative to the estimate from a simple random sample when $\rho = 1$. We compared estimates from the convenience sample when sampling weights were omitted (blue), when only the propensity score model was weighted (green), only the outcome model was weighted (purple), and when both the propensity score and outcome models were weighted (orange). Treatment effect estimates that use the true propensity scores instead of the estimated propensity scores are shown with dotted lines. Setting $\rho = 1$ implies that $K = X_1$, and since X_1 is measured we have full information on the sampling bias. We are able to obtain unbiased estimates of the sampling weights which is similar to the scenario considered by Ridgeway et al. [93] because the sampling weights contain all of the information about the selection mechanism.

Overall, weighting the outcome model has the largest impact on reducing bias but weighting the propensity score model removes additional bias. Notice that varying δ_A, δ_3, δ_4, and δ_5 have little impact on the absolute bias. This means that the size of the treatment effect does not impact the magnitude of bias. Varying one of the confounders, δ_2, has an impact on the bias but varying δ_3 does not. The reason δ_3 does not have an impact is because X_3 is squared in the outcome model (Equation 4.13) and bias can cancel out. The estimates when failing to weight both models fluctuate when varying δ_2. Covariates related to the response, Y, but not the treatment probability (δ_4 and δ_5) also do not have an impact on sampling bias. Varying δ_{AK}, or the difference in the treatment effect for the $K = 1$ and $K = 0$ subpopulations, impacts sampling bias. When sampling weights are omitted (see the blue line) the bias is minimized when interaction between K and a is 0. For sampling bias to occur, there must be treatment effect heterogeneity for subpopulations that are disproportionately represented in a convenience sample. Failing to weight either the propensity score model or the outcome model results in a minimum bias at other values of δ_{AK} because the marginal treatment effect estimate is weighted incorrectly. The marginal treatment effect can be pushed closer or farther to the estimate from the SRS, but it is difficult to predict which way it will go.

In Figure 4.2 we presented the results from varying the coefficients in the true propensity score model defined in Equation 4.12. We varied α_{02} the exponentiated coefficient on X_2 along the x-axis and α_{12} the exponentiated coefficient on $K \times X_2$ along the y-axis. We present the absolute bias relative to the estimate from the SRS when using a convenience sample and only weighting the outcome model ($\hat{\beta}_1^*$) and weighting both the propensity score model and outcome model ($\hat{\beta}_1$) in the left and right figures, respectively. Notice that the bias from the doubly-weighted model (right side) is generally stable regardless of the parameters in the propensity score model. The bias from the estimate with only the outcome model weighted (left side) is more erratic. At the given parameters of the outcome model, the bias when only weighting the outcome model is always larger than when weighting both models but it can actually be pushed to other directions when ρ decreases and δ_{AK} changes. See Figure 4.5 for an example of this. The analogous plots for varying α_{03} and α_{13} are shown in Figure 4.3. The trends are similar but less extreme because the X_3 term in the outcome model (Equation 4.13) is squared so bias of a different sign cancels. Overall, the bias when only weighting the outcome model is hard to predict and the bias when weighting both models is

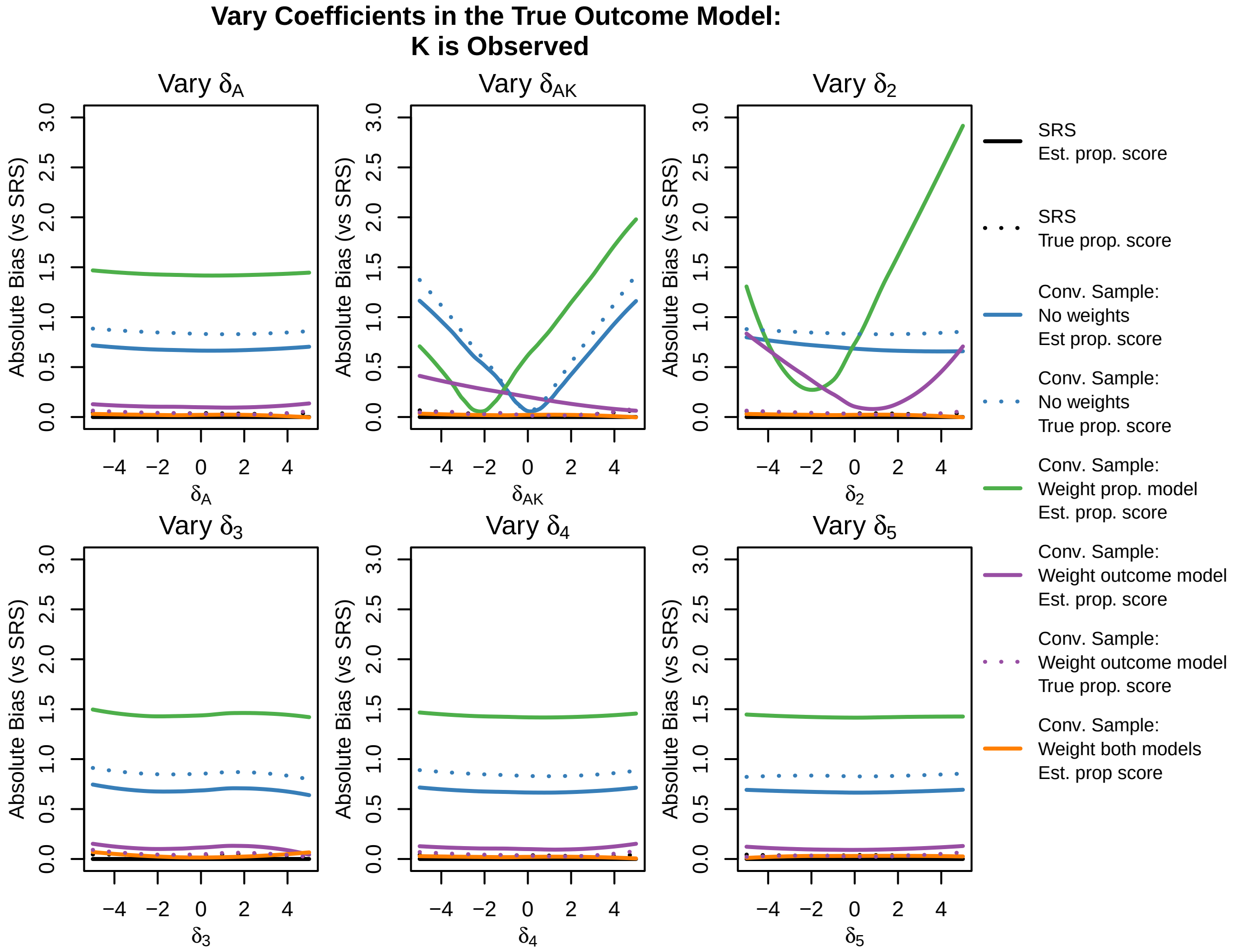


Figure 4.1: Comparison of the absolute bias of estimates of the treatment effect (β_1) relative to the estimate from a simple random sample (SRS) when different weighting strategies are used. In all figures, $\rho = 1$ and the data generating mechanism was modified by varying the coefficients in Equation 4.13.

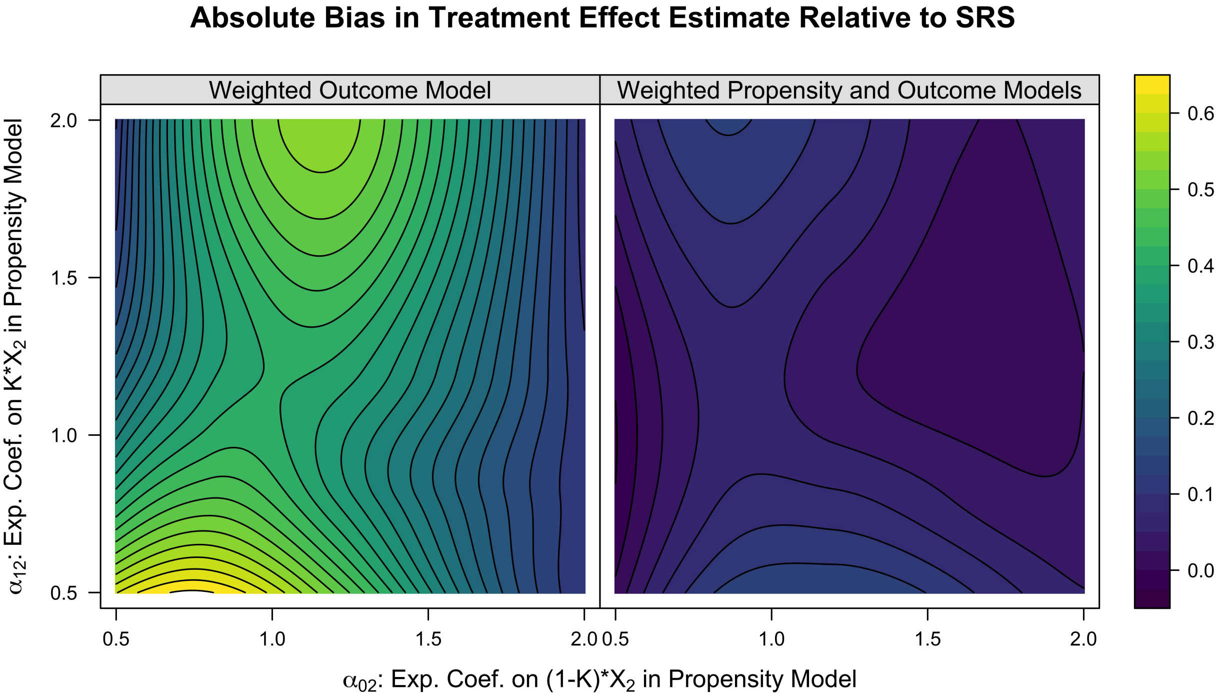


Figure 4.2: Absolute bias of treatment effect estimates ($\hat{\beta}_1$) relative to the estimate from a simple random sample (SRS) with estimated propensity scores when $\rho = 1$ and varying the parameters α_{02} and α_{12} in Equation 4.12.

more stable.

Next, we assessed the impact of not measuring K directly, but instead varying the strength of a proxy for K. In Figure 4.4, we report the mean absolute relative bias of the estimated treatment effect from the convenience sample relative to the estimate from the SRS when using different weighting strategies and varying the correlation between K^* and X_1^*. The absolute relative bias is calculated as the absolute value of the difference between the estimates from the convenience sample and the SRS divided by the bias from the SRS estimate. When $\rho = 0$, we have no information on the sampling bias, but when $\rho = 1$ we have measured everything related to the sampling bias. As ρ increases and we have more information on the sampling bias, we observe less bias in the estimated treatment effect when weighting the outcome model (orange and purple). Weighting both the propensity score model and the outcome model (orange) leads to a greater reduction in bias than only weighting the

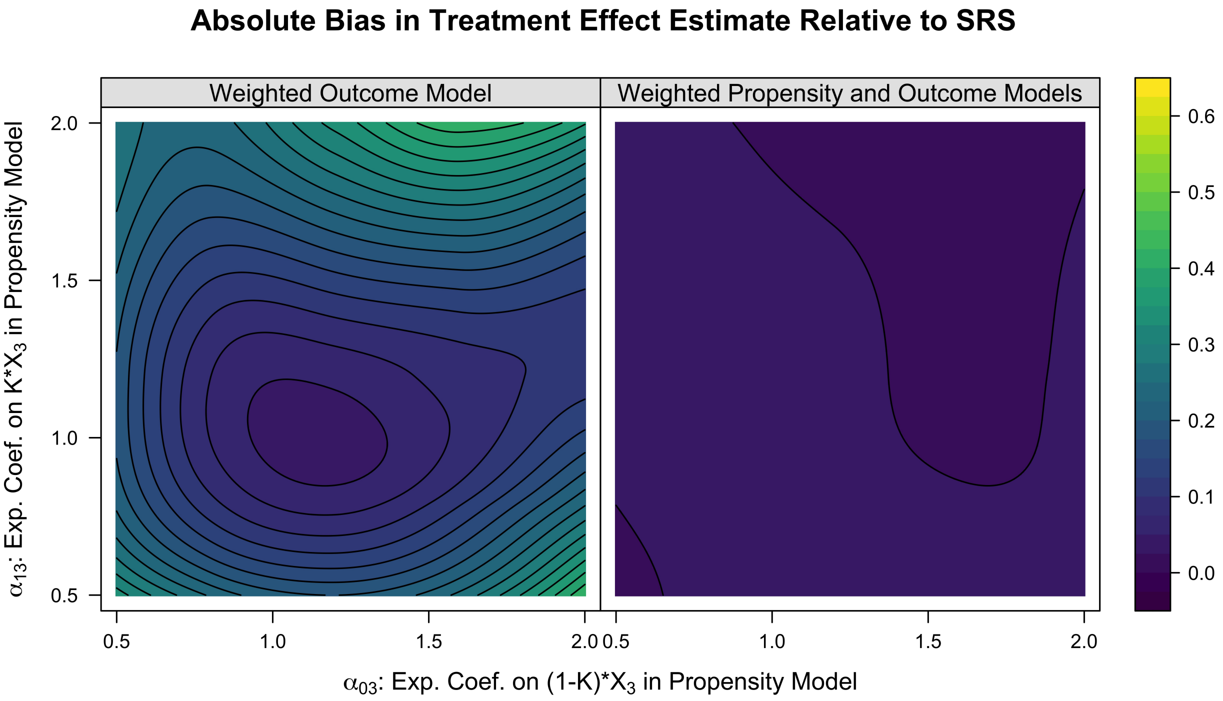


Figure 4.3: Absolute bias of treatment effect estimates ($\hat{\beta}_1$) relative to the estimate from a simple random sample (SRS) with estimated propensity scores when $\rho = 1$ and varying the parameters α_{03} and α_{13} in Equation 4.12.

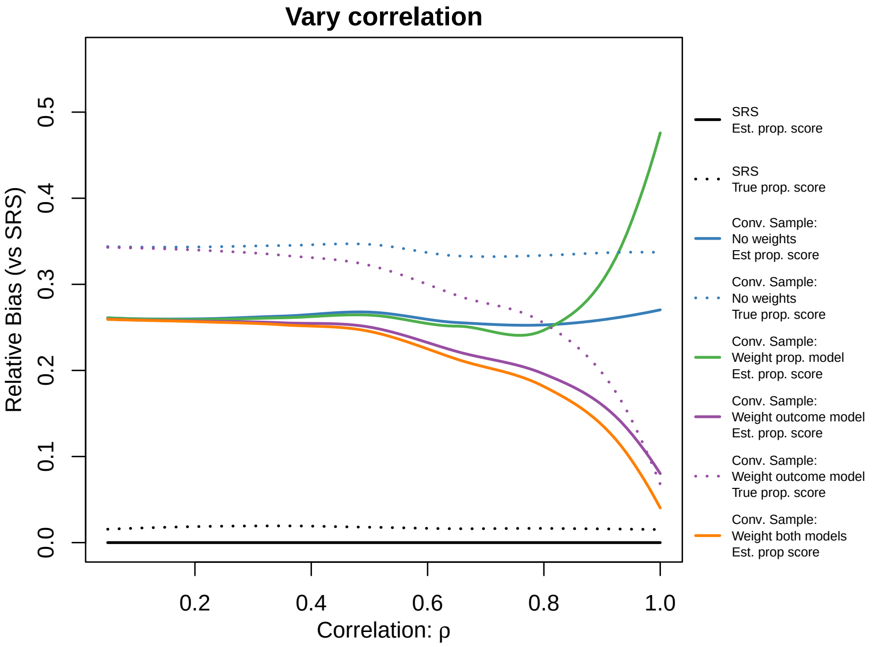


Figure 4.4: Absolute relative bias of treatment effect estimates ($\hat{\beta}_1$) relative to the estimate from a simple random sample (SRS) with estimated propensity scores when varying the correlation between K^* and X_1^*.

outcome model in this scenario. When the correlation is around 0.9, we are able to remove about half of the relative bias when weighting both the propensity score model and outcome model. A correlation of 0.9 between the continuous covariates K^* and X_1^* translates to a misclassification rate of 14% for the binary variables, where 14% of the observations of X_1 do not match K. When only weighting the propensity model (green) the relative bias increases as ρ approaches 1. This observation is consistent with the observation from Figure 4.1 that only weighting one model can lead to unexpected fluctuations in the bias. The correlation, ρ, had to be high before the sampling weights helped address sampling bias. The more information we are able to collect about the selection mechanism and the better we are able to estimate sampling weights, the better we are able to address sampling bias by weighting both the propensity and outcome models.

To illustrate how failing to weight both models can push the treatment effect estimate either direction, we repeated the top middle panel of Figure 4.1 which looks at the absolute bias of

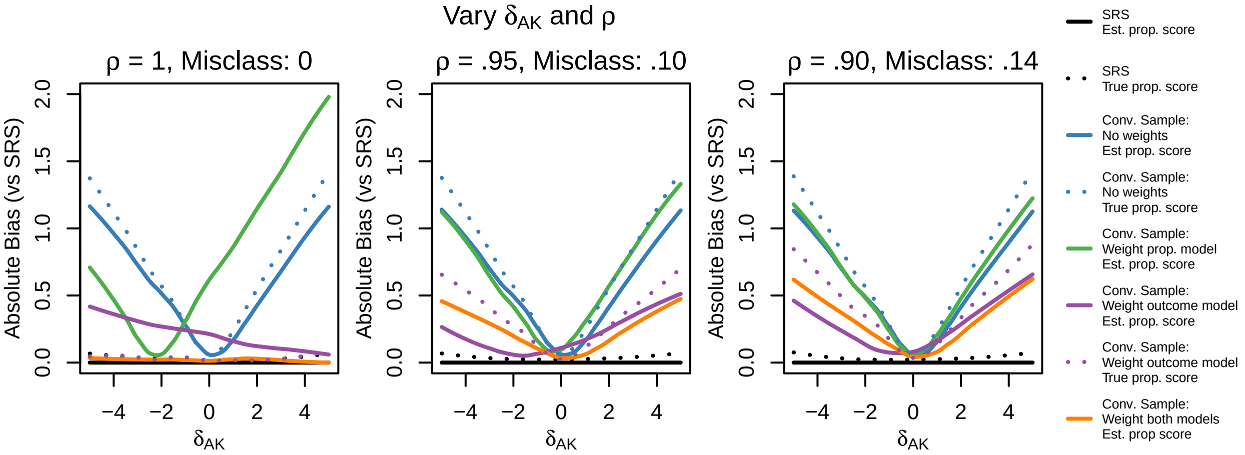


Figure 4.5: Comparison of the absolute bias of estimates of the treatment effect (β_1) relative to the estimate from a simple random sample (SRS) for different weighting strategies and varying ρ and δ_{AK}.

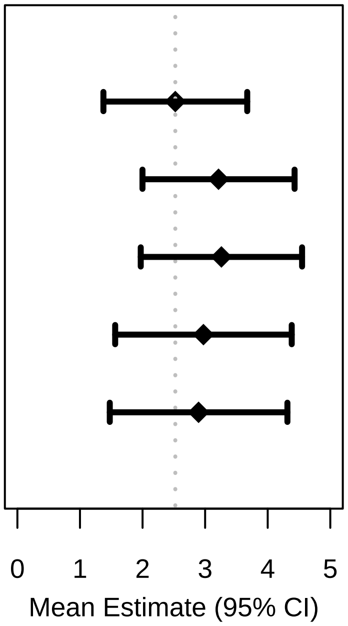
the estimated treatment effect as the magnitude of the treatment effect heterogeneity δ_{AK} varies but vary the correlation ρ in Figure 4.5. Notice that when $\rho < 0$ and $\delta_{AK} < 0$, the estimate from only weighting the outcome model (purple) has less bias than weighting both models (orange). Although one may think it would be better to only choose to weight the outcome model in some scenarios, we were unable to predict when it would have less bias than weighting both models. As we showed in Figure 4.2, only weighting the outcome model leads to volatile results. When failing to weight the propensity score model, the propensity score model is misspecified relative to the propensity score that is estimated in a SRS and leads to untrustworthy treatment effect estimates. Additionally, we see that the absolute bias when weighting both models tends to decrease as the correlation increases and the misclassification rate of X_1 increases.

Finally, we quantified the impact of accounting for the selection mechanism on the variance of the treatment effect estimates and presented the performance of our proposed variance estimate in Figure 4.6. We fixed all parameters in the data generating mechanism to the values specified in Section 4.3.1. In the table and figure on the left, we present the average estimated treatment effects and corresponding 95% confidence intervals based on the empirical standard errors across simulations. We compared estimates from a simple random sample and the convenience sample while using different weighting strategies. We observed that weighting both the propensity score model and outcome model leads to the treatment effect estimate closest to the estimate from the SRS. Additionally, the variance increased when both models were weighted which is expected because sampling weights often increase uncertainty. This increased uncertainty is a reflection of using a convenience sample and extrapolating to a target population[75, 64, 15] and reflects the true variability of the treatment effect estimation for a target population. On the right table, we compare the proposed standard error (SE) estimates, the design based SE estimate, the naive (non-sandwich) SE estimate, and the empirical SE estimate. The proposed and design based similarly which is expected since our proposed estimate is an extension to the design based estimate that adds in uncertainty from estimating sampling weights. These two estimates are similar to the empirical standard error estimate. The naive standard error estimate that does not account for including sampling weights does not come close to the empirical estimate.

4.4 Application

A recent meta-analysis summarized the evidence about the effect of vitamin E on symptoms of Alzheimer's Disease (AD) [22]. A randomized study of cognitively normal male older adults did not find an association between vitamin E supplementation and cognitive decline as measured by the Memory Impairment Screen and Consortium to Establish a Registry for Alzheimer's Disease Neuropsychological Assessment Battery [66]. A randomized study of adults with mild-to-moderate AD found a decrease in functional decline among patients assigned to a vitamin E supplement[33] and a second randomized study showed vitamin E supplementation decreased progression of AD [101]. Two additional randomized studies, one in patients with mild cognitive impairment and one in patients with AD, did not find a pos-

Average Estimated Treatment Effect (95% CI)	
Simple Random Sample	
	2.52 (1.37, 3.67)
Convenience Sample	
No weights	3.21 (2.00, 4.43)
Weight prop. model	3.26 (1.97, 4.55)
Weight outcome model	2.97 (1.56, 4.38)
Weight both models	2.90 (1.48, 4.32)



SE Estimate for Estimate With Both Models Weighted	
Empirical	0.72
Proposed	0.72
Design Based	0.72
Naive	0.62

Figure 4.6: The table and figure on the left side show the mean estimated treatment effect estimated using a simple random sample (SRS) and a convenience sample with different weighting strategies. The 95% confidence interval (CI) is constructed using the empirical standard error estimate. The table on the right compares standard error (SE) estimates of the treatment effect when when the propensity score and outcome models were weighted. The four standard error estimates are the mean proposed estimate, the mean design based estimate, the mean naive (non-sandwich) estimate, and the empirical estimate calculated across simulations.

itive effect of vitamin E [72, 90]. Three observational studies found a significant association between vitamin E supplementation and risk of progressing to AD for cognitively normal older adults [83, 118, 12], but three additional observational studies in the same population did not find this association [79, 74, 48]. We wanted to estimate the effect of vitamin E on a functional outcome with an observational study in a convenience sample and demonstrate the impact of accounting for the selection mechanism. The National Alzheimer's Coordinating Center (NACC) Uniform Data Set can be used to estimate the effects of potential treatment for Alzheimer's disease [13]. NACC data is, however, subject to sampling bias and overrepresents highly educated and non-Hispanic White participants, relative to the US population of older adults. We estimated the effect of vitamin E supplementation on the Functional Activities Questionnaire (FAQ) using propensity scores and estimated sampling weights. We compared the estimates when the propensity score estimation model was weighted, the causal model was weighted, when both models were weighted, or when neither were weighted.

4.4.1 Methods for Application

NACC data were contributed by the NIA-funded Alzheimer's Disease Core Centers and Alzheimer's Disease Research Centers (ADCs). The NACC database is funded by NIA/NIH Grant U24 AG072122. We used baseline data for NACC participants from 41 ADCs who enrolled from September 2005 to August 2021, were cognitively normal, and 65 or older. We created an indicator for prescription drug use by reviewing patients' medications listed from the last two weeks and removing common over the counter medications (such as pain pills) and vitamins. Patients who reported other medications were classified as taking prescription medications. We constructed the indicator for vitamin E supplementation by identifying individuals who reported any medication usage in the last two weeks containing the phrase ”vitamin E” as those who used a vitamin E supplement.

We estimated the probability of self-selection into NACC relative to a representative sample of adults over 65 years old from the 2013-2016 cycles of the National Health and Nutrition Examination Survey (NHANES) which we denote NHANES-REP. We use logistic regression with forward-selection using AIC for model selection and included quadratic terms in the scope as described in Chapter 3 [15]. To estimate the sampling probablities, we matched on age, sex, education, race and ethnicity, presence of high blood pressure, presence of diabetes, presence of congestive heart failure, presence of major depression, and prescription drug use. We used the estimated sampling probabilities to compute estimated sampling weights for the NACC participants. We report summary statistics from NHANES and NACC along with weighted estimates from NACC.

We summarized covariates collected in NACC by mean and standard deviation for continuous covariates and proportion for categorical variables. We compared the raw summary statistics collected in NACC and weighted summary statistics that use estimated sampling weights to those estimated with NHANES-REP. We estimated propensity scores using logistic regression with forward-selection using dAIC and AIC depending on whether the logistic model was weighted or not. We included age, sex, education, race and ethnicity, presence of high blood pressure, presence of diabetes, presence of congestive heart failure, presence of major depression, prescription drug use, type of residence, study partner type, marital status, level of independence, family history of cognitive impairment, history of smoking, thyroid disease, previous stroke, previous heart attack, and previous seizure along with quadratic terms and two-way interactions in the model scope. We estimated propensity scores by both including and omitting the estimated sampling weights. Finally, we estimated the effect of vitamin E on the FAQ score using propensity score adjustment. We estimated the effect using the unweighted and weighted propensity score estimates and compared results when including and omitting estimating sampling weights in the causal model. When both the propensity score and outcome models were weighted, we report the proposed standard error estimates detailed in Section 4.2.3. For all other estimates, we use the standard sandwich estimate that accounts for sampling weights [77]. Code for this section is available at `https://github.com/oliviabern/weighted_propscores`.

4.4.2 Results

After excluding 809 participants missing the FAQ, there were 14,358 participants available for our analysis. Unweighted and weighted summary statistics for NACC compared to those from NHANES-REP are reported in Table 4.1. When failing to account for the selection mechanism, NACC participants where more likely to be college educated (63% vs. 29%) than the general population of older adults in the US as estimated with NHANES-REP. NACC participants were also more likely to be female and have lower rates of high blood pressure, diabetes, and congestive heart failure. NACC underrepresents Hispanic and non-Hispanic Asian participants but overrepresents non-Hispanic Black participants relative to the population of older adults in the US. This is likely driven by individual Alzheimer's Disease Research Centers that focus on recruiting Black participants. Once sampling weights are incorporated into the estimation of summary statistics from NACC, they are very similar to those calculated using NHANES-REP.

When weighting both the propensity score and outcome models to obtain inference about the US population, we estimated that participants who take vitamin E supplements have a 0.29 point lower (95% CI: -0.54, -0.04) FAQ score, on average, compared to participants who do not. When failing to account for sampling bias and omitting sampling weights, we estimated that they have a 0.01 point lower (95% CI: -0.11, 0.09) FAQ score, on average. Full results are reported in Figure 4.7. The confidence intervals for the weighted methods are wider, but that reflects are increased variability from generalizing to a population different than the one sampled.

Table 4.1: Covariates are summarized as mean (standard deviation) for continuous variables and as proportions for categorical variables from samples of adults older than 65 from NHANES-REP (23,468 observations), the unweighted NACC (14,358 subjects) data set, and weighted NACC data sets. The standardized mean difference for summary characteristics in NACC relative to NHANES-REP is reported in italics and brackets. Sampling weights for NACC participants are estimated using logistic regression with forward model selection based on the AIC.

	NHANES-REP	NACC	
	Unweighted	Unweighted	Weighted
Age	69.6 (6.7)	72.0 (6.2) [-*0.34*]	69.9 (6.8) [-*0.04*]
Education: Less than high school	0.16	0.04 [*0.33*]	0.16 [-*0.01*]
Education: High school	0.23	0.14 [*0.22*]	0.23 [*0.00*]
Education: Some college	0.32	0.19 [*0.28*]	0.32 [-*0.01*]
Education: College or higher	0.29	0.63 [-*0.75*]	0.29 [*0.01*]
Race/Ethnicity: Non-Hispanic White	0.75	0.74 [*0.02*]	0.72 [*0.06*]
Race/Ethnicity: Hispanic	0.09	0.06 [*0.09*]	0.10 [-*0.04*]
Race/Ethnicity: Non-Hispanic Asian	0.04	0.03 [*0.08*]	0.05 [-*0.03*]
Race/Ethnicity: Non-Hispanic Black	0.09	0.16 [-*0.21*]	0.10 [*0.00*]
Race/Ethnicity: Other	0.02	0.01 [*0.10*]	0.03 [-*0.07*]
Female	0.54	0.65 [-*0.22*]	0.55 [-*0.01*]
High blood pressure	0.59	0.52 [*0.15*]	0.61 [-*0.04*]
Diabetes	0.22	0.13 [*0.23*]	0.24 [-*0.04*]
Congestive heart failure	0.07	0.02 [*0.19*]	0.08 [-*0.04*]
Major depression	0.10	0.09 [*0.02*]	0.08 [*0.05*]
Prescription medications	0.86	0.90 [-*0.12*]	0.86 [-*0.02*]

Higher FAQ scores correspond to needing more help with functional activities. Accounting for the selection mechanism by including sampling weights in both the propensity score and outcome models led to a larger estimated treatment effect, where the participants assigned to vitamin E had less difficulty completing activities. An FAQ score of 1 corresponds to a patient having difficulty on one activity but still being able to complete it independently [81]. The treatment effect of a third of a point is likely not clinically meaningful, but these results suggest that failing to obtain representative samples my lead to bias in the estimated treatment effect. In this application, the effect of vitamin E on FAQ scores seems to be differential for those overrepresented and underrepresented in the sample. The best way to

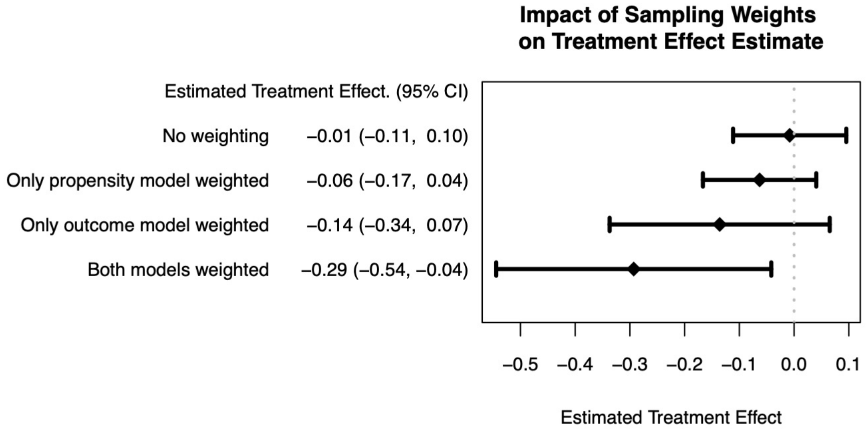


Figure 4.7: Forest plot showing the estimated treatment effect or the average difference in Functional Activities Questionnaire for older adults who take vitamin E supplements compared to those who do not. Estimates that included or omitted sampling weights in the propensity score and outcome models are compared. For the estimate that used sampling weights in the propensity score and outcome models, the reported 95% confidence intervals uses the proposed variance estimate described in Section 4.2.3, and the other three estimates use a standard robust variance estimate.

obtain representative treatment effect estimates is to recruit a representative sample, but including estimated sampling weights will allow for a more generalizable estimate.

Our results are consistent with the 3 other observations studies that have estimated an association between vitamin E supplementation and a decreased risk of cognitive decline. Using NACC data and using sampling weights to generalize to the US population considers a different patient population than clinical trials which tend to overrepresent highly educated and non-Hispanic White patients relative to the US population. Our estimates are more comparable to estimates from observational studies but those are subject to potential unobserved confounding. The NACC website (`https://naccdata.org/publish-project/authors-checklist#acknowledgment`) states "NACC data are not ideally suited to study risk factors for dementia because of varying methods of subject recruitment across Centers and because of largely incomplete exposure histories." Estimating sampling weights and using them to estimate the effect of vitamin E supplementation addresses the former limitation, but not the latter. We defined vitamin E supplementation as individuals who reported consuming vitamin E on their medications list from the previous two weeks. We may by missing identifying individuals who took vitamin E at an earlier time of their life or whose who did not include vitamin E when reporting current medications.

4.5 Discussion

In this chapter we discussed a strategy to estimate a marginal treatment effect estimate for a target population when using a convenience sample that may not be representative of that population. Bias from the selection mechanism arises when the treatment effect varies for subpopulations that are disproportionately represented in a convenience sample. Estimating sampling weights for convenience samples and using them to estimate propensity scores for a target population and corresponding propensity score adjusted treatment effect estimates for a target population is a practical solution to sampling bias. One may argue that we could directly model the treatment effect for each subpopulation that is over- or under-represented in a sample, but the proposed approach does not require us to specify which subpopulations will have treatment effect heterogeneity. In practice, we will always marginalize over some subpopulations and using sampling weights allows us to obtain a marginal estimate that generalizes to the target population. Additionally, implementing sampling weights allows us to assess the impact of sampling bias by comparing the unweighted and weighted estimated treatment effects. For example, in our application to the NACC data we estimated that vitamin E had a larger impact on functional activities scores when accounting for the selection mechanism. This suggest that subpopulations that are underrepresented in NACC have a stronger relationship between vitamin E supplementation and functional activities.

Our work extends previous papers that address combining sampling weights and propensity scores. Several papers assume the propensity score model is correctly specified and conclude that sampling weights are unnecessary in the propensity score estimation modeling[9, 68] and we agree. Ignoring the selection mechanism will not bias propensity score estimates if the model is correctly specified. We address the scenario where we are unable to correctly estimate propensity scores. Our work is most related to the analysis of Ridgeway et al. [93] and their data generation scenario 5 where the propensity score model is misspecified. Ridgeway also assumes that there is treatment effect heterogeneity with respect to a covariate that is misrepresented in the sample (see Appendix A.1 for the simulation details).

We extended this work by quantifying the impact of different features of the data generating mechanism on the sampling bias. We found that the the magnitude of treatment effect heterogeneity impacted the sampling bias–larger heterogeneity led to more bias. When looking at covariates that do not modify the treatment effect, the strength of the relationship between the covariate and the response did not change the magnitude of the sampling bias. The relationship between confounders and the probability of treatment as well as the relationship between confounders and the response impacted the volatility of estimates when only weighting the outcome model. Additionally, we expanded Ridgeway's work by considering the case when sampling weights are unknown and must be estimated, as is the case in convenience samples. We assessed the relationship between how well sampling weights are estimated and the amount of bias reduction. We considered the scenario where we collect proxies for the group membership that is overrepresented in the convenience sample and found that as the strength of the proxy increases, the bias from unrepresentative sampling decreases. We found that the correlation between the proxy and the group membership indicator had to be large before implementing sampling weights removed sampling bias. In particular, we found that when there was a 14% misclassification rate between the proxy and the covariate that determined the sampling probability, weighting the propensity score and outcome models fit in a convenience sample removed half of the sampling bias. In practice, is very important to carefully consider which covariates should be related to the sampling probability, ensure they are collected, and include them when estimating the sampling weights. Even if we are only able to collect partial information about the sampling probability, we are still able to remove some bias.

Third, we assessed the impact of accounting for the selection mechanism on the uncertainty of the treatment effect estimate and proposed a novel estimate of the variance to quantify the uncertainty. We observed that including sampling weights increased uncertainty in the treatment effect estimates which is a result of using a convenience sample to estimate a treatment effect for a different population [75, 64]. We observed this same trend in our previous work on estimating sampling weights for convenience samples [15]. Additionally, our proposed variance estimate accounts for uncertainty arising from estimating sampling probabilities and propensity scores. We observed that the proposed variance estimate matches the empirical standard errors in our simulation scenario. Our proposed variance estimate is easy to implement and so we recommend using it to account for uncertainty that arises from estimating sampling weights.

Our approach to estimating causal effects in a convenience sample with observational data does have limitations. As in all observational studies, the treatment effect estimate can be subject to unmeasured confounding if the unconfoundedness assumption fails. Using propensity scores to adjust for observed confounding, however, will allow us to better reduce bias in estimated causal effects than ignoring all confounding. It is important to carefully consider potential confounders and attempt to collect them. Similarly, we are only able to account for measured sampling bias. Sampling bias occurs when the treatment effect is differential for subpopulations misrepresented in the sample. Researchers should hypothesize which variables are likely related to the sampling probability and may cause treatment effect modificiation. In clinical studies, treatment effect may be differential by socioeconomic status, health status, sex, education, race, or ethnicity. It is important to collect the hypothesized variables in the convenience sample and use a representative sample that also collects these variables. NHANES is a practical option in biomedical application because it contains information on group identifiers that may be under- or over-represented in your sample. As we saw in our simulation study, only collecting partial information on the selection mechanism still reduced bias relative to estimates from a simple random sample. Thus, adjusting for measured sampling bias is better than ignoring it. In our application, we are unable to quantify how well we are able to estimate sampling weights for NACC participants, but we were able to account for sampling bias related to race, ethnicity, education level, sex, age, and several comorbidities. We are unable to account for sampling bias caused by differences in socioeconomic status or unmeasured health concerns. Both weighted estimates and unweighted estimates have the same direction of effect, but weighted estimates shift further away from the null hypothesis. Incorporating estimated sampling weights leads to increased variability in estimates but that is preferable to ignoring the selection mechanism and providing a biased treatment effect estimate for the population of interest. Directly collecting a representative sample is the best solution to avoid sampling bias, but it may not always be practical or ethical. Estimating sampling weights allows us to better estimate treatment effect estimates in a target population.

Oversampling individuals from certain subpopulations with differential treatment effects can lead to biased estimates of the treatment effect in a target population. Estimating sampling weights for a convenience sample allows us to address and quantify the impact of sampling bias. Using our proposed variance estimate allows for quantification of the treatment effect estimate in a target population.

Chapter 5

Comparison of analytic and resampling estimates of prediction error under a biased sampling scheme

5.1 Introduction

In the previous chapter, we discussed methods for incorporating sampling weights into propensity scores with convenience samples and in this chapter we will discuss methods for incorporating sampling weights into predictive models for convenience samples. Prediction models are common in healthcare and many other fields including political science, physics, technology, finance, and biology. Although accurate predictive models are beneficial, inaccurate predictions can be misleading at best and unethical at worst. A common way to select a prediction model from a set of candidate models is to select the model that minimizes the prediction error in the population of interest [54]. If the training sample is not representative of the population, estimates of the prediction error will likely be biased for the error in the target population. Biased estimates of the prediction error can result in failing to select the best predictive model leading to inaccurate predictions.

The Consent-to-Contact (C2C) registry at the University of California (UC), Irvine enrolls volunteers who are interested in participating in future research studies [51]. C2C researchers want to learn about recruitment strategies for the United States and want to gain more information about strategies for neighborhoods with greater socioeconomic disadvantage. As part of a recently funded NIH grant, the leadership of the C2C registry plans to oversample this subpopulation. Once this is implemented, if the C2C data set is used to build predictive models the predictions will not generalize to the target population of the United States. To address this, the sampling scheme needs to be accounted for when fitting and evaluating the performance of a predictive model.

We can broadly classify methods for assessing predictive model performance into two classes: analytic methods and resampling methods. Analytic methods such as Mallows C_p [78], Akaike's information criterion (AIC) [4, 5], and Bayesian information criterion (BIC) [104] derive an analytic relationship between the training error and the prediction error to estimate the prediction error of a given model. Resampling methods such as cross-validation [106, 107, 6] and the bootstrap [35] resample the data to mimic having a separate test set for estimating the prediction error. All of these methods assume that the sample is representative of the population or that the sample is collected using a simple random sampling scheme (SRSS).

If a training sample is not representative of the target population, these estimates will not give a true assessment of a model's performance in the target population. A biased sampling scheme (BSS) is one where observations are sampled with unequal sampling probabilities, leading to an unrepresentative sample. Survey samples are one example of a BSS where researchers sample individuals with non-uniform, but pre-specified sampling probabilities. The resulting sample is not representative of the target population, but sampling weights for each observation are known because they are a function of the sampling probabilities (see chapter 2 of Lohr (2010) for an overview [73]). Convenience samples are another example of a sample with a BSS where researchers select individuals based on their availability (see Section 1.3 of Lohr (2010) [73]). For example, in a research registry the participants are people who were interested in volunteering and were informed about the opportunity. Sampling weights must be estimated for convenience samples because the sampling probabilities are not pre-specified. Sampling weights are commonly used in the survey sampling literature to obtain estimates for the target population when using a sample with non-uniform sampling probabilities [75, 59, 73]. Sampling weights estimated with a representative sample are used for convenience samples to obtain estimates for the target populations [2, 15, 26, 94, 40, 117, 88].

There are several extensions to methods for estimating the prediction error that have been developed for a BSS. These extensions use sampling weights in the estimate of the prediction error for the target population. Lumley and Scott (2015) developed an extension to the AIC called the design-based AIC (dAIC) [76]. Wieczorek et al. (2022) developed an extension to 5-fold cross-validation that implements sampling weights [115]. In Lumley and Scott's paper they derive a relationship between dAIC and weighted cross-validation and show that the two are asymptotically equivalent. No one has, however, empirically compared their performance for estimating out-of-sample predictive error. In the SRSS case, it has been shown that resampling methods are a better estimate of out-of-sample predictive error because they directly estimate it but are more computationally expensive [54]. The objective of this chapter is to evaluate the performance and trade-offs of analytic and resampling estimates of predictive error with a BSS.

The remainder of this chapter is organized as follows. We introduce common estimates of the prediction error under a SRSS in Section 5.2 and under a BSS in Section 5.3. We discuss the theoretical relationship between analytic and resampling methods developed by Lumley and Scott (2015) [76] in Section 5.4. Next, we present a simulation study comparing analytic and resampling methods for estimating the prediction error under a BSS in Section 5.5. In Section 5.6 we then apply these methods to real data and compare out-of-sample predictive estimates for models designed to predict individual willingness to participate in research using a sub-sample of C2C data with larger sampling probabilities for individuals from neighborhoods with more socioeconomic disadvantage in. We conclude with a discussion of the considerations for deciding between analytic and resampling methods when data are obtained via a BSS in Section 5.7.

5.2 Prediction assessment methods under a simple random sampling scheme

For a more comprehensive overview of prediction assessment see Chapter 7 of Hastie, Tibshirani, and Freedman (2009) [54]. Briefly, a common strategy for selecting a prediction model from a scope of models is to choose the model with the lowest prediction error in the target population. We quantify the prediction error using a loss function such as $0-1$ loss, squared error loss, or log-likelihood loss. A loss function takes the form, $L(Y, \widehat{f}(X))$ where X is a vector of covariates, Y is the response, and $\widehat{f}(X)$ is the predicted value of Y using model $\widehat{f}(\cdot)$. The log-likelihood loss is

$$L(Y, \widehat{f}(X)) = -2 \times \text{loglik} = -2 \times \ell\big(\widehat{f}(X)\big) = -2 \times \log \Pr(Y|X)$$

where loglik and $\ell()$ both denote the log-likelihood. Note that the -2 term out front is included so log-likelihood loss is equivalent to squared error loss when a normal likelihood is assumed, because for relative model comparison the scaling factor is inconsequential. A common analytic method for estimating prediction error, Akaike's information criterion (AIC), uses the log-likelihood loss and so in this chapter we focus on the log-likelihood loss for ease of comparison, but all of the methods discussed can be adapted to other loss functions.

The prediction error, as measured by the loss function, in the target population is formalized by the expected prediction error (EPE) or expected test error for an independent sample drawn from the target population. The EPE under log-likelihood loss is

$$\text{EPE} = E_{(X^0,Y^0)}\Big[L(Y^0, \widehat{f}(X^0))\Big] = -2E_{(X^0,Y^0)}\Big[-2 \times \ell\big(\widehat{f}(X)\big)\Big] \tag{5.1}$$

where the expectation is over the joint distribution of X and Y. The notation X^0 and Y^0 denote new values of X and Y drawn from their joint distribution. If the data set is large enough one can partition it into separate training and test sets. For smaller data sets the full sample is commonly used for model training, selection, and assessment.

A naive estimate of the EPE is the loss function computed with the training sample which is termed the *training error*. It is well known that training error underestimates the true EPE because the model is fit and evaluated with the same data. Suppose we have a training sample with n independent observations $\{(x_i, y_i)\}$ where $i = 1, \ldots, n$ that we use to estimate $\widehat{f}(x_i)$. Let

$$\ell\big(\widehat{f}(X)\big) = \sum_{i=1}^{n} \ell_i\big(\widehat{f}(x_i)\big)$$

where $\ell_i\big(\widehat{f}(x_i)\big)$ is the i-th observation's contribution to the log-likelihood. Then the training error (err) under log-likelihood loss is

$$\text{err} = \frac{1}{n}\sum_{i=1}^{n} L(y_i, \widehat{f}(x_i)) = -2 \times \sum_{i=1}^{n} \ell_i\big(\widehat{f}(x_i)\big).$$

There are two classes of estimators of prediction error using training data. The first class consists of analytic methods that estimate the difference between the test error and the training error. The second class, resampling methods, resample the data to mimic having separate training and test sets. There have been many estimates in both classes developed that assume the training set is representative of the target population. We will compare several examples of analytic and resampling methods, but there are many options to choose from for different loss functions and estimation strategies.

5.2.1 Analytic estimates for a SRSS

For ease of computation, analytic methods estimate the prediction error when the covariate distribution of X is held constant. This can be interpreted as the expected prediction error if n new values of the response, $y_i^0, i = 1, \ldots, n$ were drawn for each x_i value in the training set. This is called the *in-sample error* ($\mathrm{Err_{in}}$). Under log-likelihood loss $\mathrm{Err_{in}}$ is given by

$$\mathrm{Err_{in}} = \frac{1}{n}\sum_{i=1}^{n} E_{Y^0}\Big[L(Y_i^0, \widehat{f}(x_i))\big|Y\Big] = -2\sum_{i=1}^{n} E_{Y^0}\Big[\ell_i\big(\widehat{f}(x_i)\big)\big|Y\Big].$$

Notice that the expectation is only over the distribution of Y^0. *Optimism* (op) is defined as the difference between the in-sample error and the training error,

$$\mathrm{op} \equiv \mathrm{Err_{in}} - \mathrm{err}.$$

Optimism tends to be positive since the training error generally underestimates $\mathrm{Err_{in}}$. In most cases it is easier to estimate the average optimism, ω, where the expectation is taken over Y,

$$\omega \equiv E_Y(\mathrm{op}).$$

Thus analytic estimates of the in-sample error generally take the form

$$\widehat{\mathrm{Err}}_{\mathrm{in}} = \mathrm{err} + \widehat{\omega}.$$

Akaike's information criterion (AIC)

Akaike's information criterion (AIC) [4, 5] is an analytic estimate of the in-sample error under log-likelihood loss that follows the above form. Let $\widehat{\theta}$ be the maximum likelihood estimates of θ, a vector of parameters defining prediction model $f(\cdot)$. Then $\mathrm{Err}_{\mathrm{in}}$ under log-likelihood loss is estimated by

$$\mathrm{AIC} = -2\ell\big(\widehat{f}(X)\big) + 2p. \tag{5.2}$$

To emphasize the dependence of the log-likelihood on $\widehat{\theta}$ we could change our notation in this formula from $\ell\big(\widehat{f}(X)\big)$ to $\ell\big(\widehat{\theta}\big)$. The Bayesian information criterion (BIC) [104] is another common analytic estimator under log-likelihood loss and Efron's general covariance penalty [36] extends analytic estimates to a broader class of loss functions. In this chapter we will focus on the AIC as an example of analytic methods, but other estimators should perform similarly.

5.2.2 Resampling estimates for a SRSS

Resampling estimates are able to directly estimate the EPE because they train the model and test it on non-overlapping samples which is analogous to drawing new observations for both X and Y.

$\boldsymbol{K}$-fold cross-validation

K-fold cross-validation (CV) [106, 107, 6] mimics partitioning the data into training and test sets by splitting the data into K folds. For $k = 1, \ldots, K$ the model is fit on all folds except fold k and the loss function is computed for the k-th fold. The loss function is then aggregated across all the folds. More formally, let $k(i)$ be the fold that containing observation i and let $\widehat{f}^{-k(i)}(x_i)$ denote the predicted value for y_i from the model fit on all folds besides $k(i)$. Then the K-fold CV statistic under log-likelihood loss is

$$\mathrm{CV}_K = \frac{1}{n}\sum_{i=1}^{n} L(y_i, \widehat{f}^{-k(i)}(x_i)) = -2\sum_{i=1}^{n} \ell_i(\widehat{f}^{-k(i)}(x_i)). \tag{5.3}$$

For large values of K, the cross-validation estimate is subject to high variability because there is a high degree of overlap across the folds, but for small values of K the estimate can have high bias because the training samples are smaller. Setting $K = 5$ or 10 is often recommended as a compromise and we will use $K = 5$ in this chapter [54, 21, 65].

Leave-one-out cross-validation

When $K = n$, the CV estimator is termed the leave-one-out (LOO) CV estimator. In general, it is expensive to compute but there is an analytic form of the estimator for linear regression models. Specifically, the analytic form of the LOO statistic in the ordinary least squares model is based on the leave-one-out residuals, or the residuals that would be obtained if the model was fit without y_i [6]. Let $\widehat{y}_i = \widehat{f}(x_i)$ and $\widehat{y}_{(i)}$ be the predicted value of y_i from the model fit with all observations besides y_i. Then let $\mathbf{X}$ be the $n \times p$ matrix of p covariates where the i-th row is x_i. Then the leave-one-out residual for y_i is,

$$y_i - \widehat{y}_{(i)} = \frac{y_i - \widehat{y}_i}{1 - \mathbf{H}_{ii}}$$

where $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ is the hat matrix. The LOO CV estimator under log-likelihood loss when Y is assumed to follow a normal distribution with variance σ^2 is,

$$\text{CV}_{\text{LOO}} = -2n \log\left(\frac{1}{\sigma\sqrt{2\pi}}\right) + \sum_{i=1}^{n} \frac{1}{\sigma^2}\left[\frac{y_i - \widehat{y}_i}{1 - \mathbf{H}_{ii}}\right]^2 \tag{5.4}$$

where σ is often estimated with a low bias model.

5.2.3 .632 bootstrap

Bootstrapping is a versatile technique commonly used to estimate a sampling distribution for summary statistics [34, 37]. Suppose we have data $\mathbf{Z} = \{z_1, ..., z_n\}$ where $z_i = (x_i, y_i)$. Bootstrapping involves sampling from Z with replacement B times to obtain B bootstrap samples of size n. The general approach of a bootstrap (BS) estimate is to compute a statistic for $b = 1, \ldots, B$ bootstrap samples and observe it's behavior across samples. Bootstrapping can be used to estimate the expected prediction error, but there is overlap across bootstrap samples so they can not act as separate training and test sets as is done in cross-validation. The leave-one-out bootstrap fits the model on the b-th bootstrap sample and computes the loss function on all z_i from the full sample that were not sampled into the b-th bootstrap sample.

Let $\widehat{f}^{*b}(x_i)$ denote the predicted value for y_i from the model fit on the b-th bootstrap sample, C_{-i} be the set of indices of bootstrap samples that do not include z_i, and $|C_{-i}|$ be the size of the set. Then the leave-one-out bootstrap estimate of the EPE under log-likelihood loss is

$$\text{BS}_{\text{LOO}} = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{|C_{-i}|} \sum_{b \in C_{-i}} L(y_i, \widehat{f}^{*b}(x_i)) = -2\sum_{i=1}^{n} \frac{1}{|C_{-i}|} \sum_{b \in C_{-i}} \ell_i(\widehat{f}^{*b}(x_i)). \tag{5.5}$$

The LOO BS estimator is subject to high bias from small sample sizes because each sample is drawn with replacement. The probability a given observation will be sampled into the b-th bootstrap sample is

$$\Pr(z_i \in \text{ BS sample } b) = 1 - \left(1 - \frac{1}{n}\right)^n \approx 1 - e^{-1} = 0.632.$$

The .632 bootstrap estimator [35] is a weighted average between the LOO BS (which overestimates the EPE) and the training error (which underestimates the EPE). The weights are based on the probability of sample inclusion. The .632 BS estimate of the EPE is

$$\text{BS}_{.632} = .368 \cdot \text{err} + .632 \cdot \text{BS}_{\text{LOO}}. \tag{5.6}$$

This estimator does not perform well for very overfit models.

5.3 Prediction assessment methods for a biased sampling scheme

All of the prediction assessment methods discussed above assume the training sample is representative of the target population. Samples obtained via a BSS are, however, not representative of the target population. Directly using the above estimates of the EPE will provide estimates of the prediction error for the population one's sample is representative of. Several methods have been proposed for incorporating sampling weights that have been developed to estimate the prediction error in a target population for a model trained on data from a BSS. In this section we assume that all prediction models are fit using sampling weights to account for the sampling scheme. For example, suppose a weighted linear regression model is fit. The goal of weighted prediction assessment methods is to estimate how well the estimated prediction model will perform in the target population.

5.3.1 Analytic estimates for a BSS

Similar to analytic estimates for the EPE for a SRSS, analytic estimates that account for the sampling scheme target the in-sample error by estimating the expected optimism. Sampling weights are included in the computation of the training error and the optimism.

Design-based AIC

Lumley and Scott (2015) extended AIC (Equation 5.2) to include sampling weights to account for the sampling scheme [76]. Let w_i be the sampling weight for observation i where $\sum_{i=1}^{n} w_i = 1$. The weighted estimate of the training error under log-likelihood loss is

$$\bar{\ell}(\theta) = \sum_{i=1}^{n} w_i \ell\big(\widehat{f}_\theta(x_i)\big) = \sum_{i=1}^{n} w_i \ell_i(\theta).$$

Then $\bar{\theta}$ is the value of θ that maximizes $\bar{\ell}(\theta)$. Suppose the maximum of the expected population log-likelihood $\ell(\theta)$ is obtained at a unique point θ^*. Under some regularity conditions $\bar{\theta}$ is consistent for θ^* as $n, N \to \infty$ where N is the population size [76, 42]. We also assume the asymptotic framework that there is a sequence of finite populations that are random samples from the target super-population.

A weighted estimate of the score function that accounts for the sampling scheme is

$$\bar{U}(\theta) = \frac{\partial \bar{\ell}(\theta)}{\partial \theta}$$

where $\bar{\theta}$ solves the corresponding score equation $\bar{U}(\bar{\theta}) = 0$. Let $\bar{\mathcal{I}}(\bar{\theta})$ be the weighted analogue of Fisher's observed information matrix,

$$\bar{\mathcal{I}}(\theta) = -\frac{\partial \bar{U}(\theta)}{\partial \theta} = -\sum_{i=1}^{n} w_i \frac{\partial^2 \ell_i(\theta)}{\partial \theta \partial \theta^T}.$$

The asymptotic covariance of $\sqrt{n}\bar{\theta}$ can be estimated with a sandwich estimate:

$$\bar{V}(\bar{\theta}) = \bar{\mathcal{I}}(\bar{\theta})^{-1}\bar{V}_U(\bar{\theta})\bar{\mathcal{I}}(\bar{\theta})^{-1}$$

where $\bar{V}_U(\theta)$ is a consistent estimator of the covariance of $\sqrt{n}\bar{U}(\theta)$ such as a method of moments estimator. The design-based AIC, or dAIC, that accounts for the sampling scheme is given by

$$d\text{AIC} = -2\bar{\ell}\big(\bar{\theta}\big) + 2\text{trace}(\bar{\mathcal{I}}(\theta)\bar{V}(\bar{\theta})). \tag{5.7}$$

If the weights are uniform and the model is correctly specified this reduces to the usual AIC (Equation 5.2). There are also survey weighted extensions to the BIC [76] and Efron's general covariance penalty [57].

5.3.2 Resampling estimates for a BSS

Since resampling estimates directly estimate the EPE using mutually exclusive training and test sets, they do not require estimating (and accounting for the sampling scheme in) the optimism. Instead, the sampling scheme can be accounted for by including sampling weights in the computation of the loss function on the test sets.

Survey weighted K-fold cross-validation

An extension of the K-fold cross-validation estimate (Equation 5.3) for a BSS was recently proposed by Wieczorek et al. (2022) [115]. Under log-likelihood loss the weighted K-fold

CV statistic is given by

$$wCV_K = \frac{1}{n}\sum_{i=1}^{n} w_i L(y_i, \widehat{f}^{-k(i)}(x_i)) = -2\sum_{i=1}^{n} w_i \ell_i(\widehat{f}^{-k(i)}(x_i)).$$

Notice the weights scale each observation's contribution to the loss function.

Survey weighted leave-one-out cross-validation

Similarly, there is a natural weighted extension to the leave-one-out CV estimator (Equation 5.4) for weighted linear regression under a BSS. Assuming a log-likelihood loss function and that Y follows a normal distribution with variance σ^2, the weighted LOO CV estimate is

$$wCV_{\text{LOO}} = -2n\log\left(\frac{1}{\sigma\sqrt{2\pi}}\right) + \sum_{i=1}^{n} w_i \frac{1}{\sigma^2}\left[\frac{y_i - \bar{y}_i}{1 - \bar{\mathbf{H}}_{ii}}\right]^2$$

where $\bar{y}_i$ is the estimate from solving the weighted score equation $\bar{U}(\theta)$. In the scenario under consideration where the model is fit with weighted least squares, if $\mathbf{W}$ is the $n \times n$ diagonal matrix of weights w_i then $\bar{\mathbf{H}} = \mathbf{X}(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}$ and $\bar{y}_i = \bar{\mathbf{H}}_{ii}y_i$.

Cawley (2006) proposed a weighted LOO CV estimator for weighted linear regression support vector machines with a categorical response but they discuss weights for balancing the response categories to match the population prevalence [24].

Survey weighted .632 bootstrap

A weighted extension to the leave-one-out BS estimator (Equation 5.5) for a BSS with log-likelihood loss is

$$wBS_{\text{LOO}} = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{|C_{-i}|}\sum_{b\in C_{-i}} w_i L(y_i, \bar{f}^{*b}(x_i)) = -2\sum_{i=1}^{n}\frac{1}{|C_{-i}|}\sum_{b\in C_{-i}} w_i \ell_i(\bar{f}^{*b}(x_i)).$$

We can denote the weighted training error as $w\mathrm{err} = \bar{\ell}(\bar{\theta})$, then the weighted extension to the .632 bootstrap estimator (Equation 5.6) is

$$w\mathrm{BS}_{.632} = .368 \cdot w\mathrm{err} + .632 \cdot w\mathrm{BS}_{\mathrm{LOO}}.$$

Rowe and Binder (2008) [96] discussed a weighted version of the leave-one-one bootstrap for stratified samples based on a bootstrap method proposed by Rao, Wu, and Yue (1992) [92]. They compared it to a weighted AIC that did not account for the sampling scheme in the optimism, as is done for dAIC, in an applied data analysis. They did not provide any theoretical justification or simulation studies to assess the performance of their method.

5.4 Asymptotic relationship between AIC and leave-one-out cross-validation

There are trade-offs of using analytic or resampling methods to estimate the EPE [54]. Resampling methods directly estimate the EPE instead of the training error, but they suffer from high computational cost. Analytic methods estimate the in-sample error which generally underestimates the EPE since it conditions on the observed distribution of the predictor space, X. They are also constrained by which loss functions are available, though Efron's general covariance penalty extended analytic methods to loss functions from the *q class* of error measures. Analytic methods are, however, easy to implement and quick to compute, so it is of interest to see how well they perform compared to resampling methods.

For a SRSS, Stone (1977) [107] showed that AIC and leave-one-out CV are asymptotically equivalent assuming the true model is considered in the scope. Lumley and Scott (2015) [76] derived a similar relationship for the dAIC and weighted leave-one-out CV in Section 3.1 and Appendix A.1 of their paper.

Briefly, the weighted LOO CV estimator of the expected population log-likelihood is

$$\bar{\ell}_{w\text{LOO}} = \bar{\ell}(\bar{\theta}_{(i)}) = \sum_{i=1}^{n} w_i \ell_i(\bar{\theta}_{(i)})$$

where $\bar{\theta}_{(i)}$ is the vector of parameters from fitting the weighted model without observation i. The LOO log-likelihood can be related to the model fit on the full data as follows,

$$\bar{\ell}_{w\text{LOO}} = \bar{\ell}(\bar{\theta}) - \text{trace}\big(\bar{\mathcal{I}}(\bar{\theta})\bar{V}_J\big) + o_p(n^{-1})$$

where $\bar{V}_J$ is a jackknife estimator of $\text{Cov}(\bar{\theta})$:

$$\bar{V}_J = \frac{n-1}{n} \sum_{i=1}^{n} \big(\bar{\theta}_{(i)} - \bar{\theta}\big)^2.$$

From Equation 5.7, the dAIC estimator of the expected population log-likelihood is,

$$\bar{\ell}_{d\text{AIC}} = \bar{\ell}(\bar{\theta}) - \text{trace}\big(\bar{\mathcal{I}}\big(\bar{\theta}\big)\bar{V}(\bar{\theta})\big).$$

Recall that $\bar{V}(\bar{\theta})$ is the asymptotic covariance of the maximum likelihood estimate of θ. If the jackknife estimate $\bar{V}_U$ is also a consistent estimate of θ, the dAIC and weighted LOO CV estimates are asymptotically equivalent.

Thus, the AIC and dAIC should approach the unweighted and weighted LOO estimators, respectively, as the sample size increases which is when resampling methods become more computationally expensive. In light of this, analytic methods should be more attractive in larger samples because they will yield better estimates of the EPE and save computational resources precisely when needed. We will empirically explore this later via a simulation study.

5.5 Simulation study

We conducted a simulation study to compare the performance of unweighted and weighted estimates of the EPE along with analytic and resampling estimates. We designed a simulation study that would reflect our application. We used the same data set from the C2C registry but simulated the response so we could control the data generating mechanism and sampling scheme. As part of the enrollment process participants answer questions about demographics, medical history, behaviors, and attitudes towards research.

The Area Deprivation Index (ADI) is a measure of socioeconomic disadvantage at the census block level [63, 109]. The ADI includes features such as income, education, employment, and housing quality. ADI ranks neighborhoods within individual states or across the US with each neighborhood being assigned to a decile. We used ADI scores based on the state rankings. Scores range from 1 to 10 where a score of 1 represents neighborhoods with the least socioeconomic disadvantage and a score of 10 representing the most socioeconomic disadvantage. To obtain more information about recruitment strategies for neighborhoods with more socioeconomic disadvantage, the future plan for the C2C registry is to recruit participants with a stratified approach so that there is equal representation across the ten levels of ADI.

To replicate this sampling scheme, we treated the C2C sample with N = 2,822 observations as a finite population and drew training samples with a BSS using sampling probabilities for each ADI score inversely proportional to the proportion of that score in C2C. An illustration of the sampling scheme and the prevalence of each ADI score in samples drawn with a SRSS and a BSS is provided in Figure 5.1. We fit weighted linear regression models to predict a simulated continuous response, Y, and compared estimates of the expected prediction error in the population. Note that all of the predictive models are weighted regardless of whether or not the prediction error estimate is weighted.

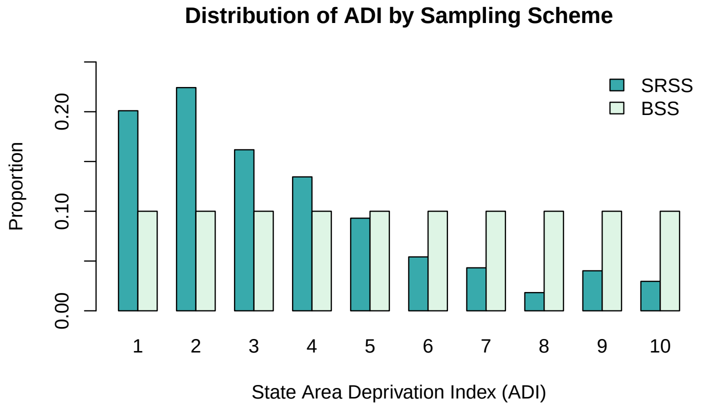


Figure 5.1: The distribution of state Area Deprivation Index (ADI) scores in samples drawn with a simple random sampling scheme (SRSS) and a biased sampling scheme (BSS).

The covariates collected on C2C participants included age, Cognitive Function Instrument score (CFI) [7, 112], Research Attitudes Questionnaire score (RAQ) [99]), years of education (educ), primary language English (English primary), enrollment after the start of the COVID-19 pandemic on 3/19/2020 (join > 3/19/20), exercise, sex, prescription medicine use (medications), reporting 0, 1, or 2+ medical conditions (N. medical cond), race and ethnicity, consenting to receive calls from C2C researchers (receive calls), consenting to receive mail from C2C researchers (receive mail), whether they learned about the C2C from an email, community talk, or other method (referral method), and signing up for the email newsletter (enewsletter). Race and ethnicity were grouped into five mutually exclusive categories. The Hispanic category includes all participants who reported a Hispanic ethnicity. The non-Hispanic (NH) Asian, NH Black, and NH White categories include participants who did not report a Hispanic ethnicity but reported only one race which were Asian, Black or African American, and White, respectively. The Other category includes all remaining participants who who reported a non-missing race including multiple races. We originally included an additional covariate, agreeing to receive emails from C2C researchers, but almost all (98.6%) of participants agreed and so we removed it from the model scope to avoid problems with sparsity.

5.5.1 Simulation methods

We conducted 250 simulations and within each simulation we drew a training sample using a BSS of size $n = 500$ and drew a test set using a SRSS of the same size. We simulated a normally distributed response, $y_i \sim \text{Normal}(\mu_i, 1)$, to reflect a continuous measure of research willingness where

$$\begin{aligned}\mu_i = 30 &+ .8(\text{Receive Calls}_i) - .5log(\text{RAQ}_i) \times (\text{ADI}_i) + .5(\text{Receive Mail}_i)\\ &+ .5(\text{Sex}_i) + .7(\text{enewsletter}_i) - .03(\text{age}_i) + .2(\text{Hispanic}_i) - .4(\text{NH Asian}_i)\\ &+ .4(\text{NH Black}_i) + .7(\text{medications}_i) \times (2\text{+ medical conditions}_i) \quad (5.8)\end{aligned}$$

We then fit weighted linear regression models in the training sample drawn with a BSS with different subsets of the covariates mentioned above as predictors. We included linear and quadratic terms for all continuous covariates in our model scope. For each model fit, we estimated the prediction error using unweighted estimates: AIC, 5-fold cross-validation (CV_5 or 5-fold CV), leave-one-out cross-validation (CV_{LOO} or LOO CV), and the .632 bootstrap ($\text{BS}_{.632}$ or .632 BS). We also estimated the prediction error with weighted estimates: the design-based AIC (dAIC), weighted 5-fold cross-validation ($w\text{CV}_5$ or 5-fold wCV), weighted leave-one-out cross-validation ($w\text{CV}_{\text{LOO}}$ or LOO wCV), and the weighted .632 bootstrap ($w\text{BS}_{.632}$ or .632 wBS). These methods are defined in Sections 5.2 and 5.3. Throughout, we assumed the sampling weights were known. For comparison, we computed the prediction error on the test set sampled with a SRSS which functions as an independent test set from the target population. We refer to this quantity as the EPE in our results.

Ideally, we would have used best subsets regression and picked the best model from the full scope of candidate models. There were, however, too many potential predictors for this to be computationally feasible and we instead used stepwise forward selection to determine the path for adding covariates. To mimic best subsets, within each simulation we determined the order covariates would be added by using forward selection with the EPE calculated on the test set. We did not stop adding covariates when the estimated loss stopped decreasing, but continued adding covariates until all covariates were included. We then estimated the prediction error with the four unweighted and four weighted methods for those same models. This allowed us to make a fair comparison across methods because we were comparing the same scope.

To assess each method for estimating the EPE's performance for model selection, we determined which model had the lowest prediction error to decide which model was chosen within each simulation. We compared how frequently each covariate was chosen as part of the most optimal model by each EPE estimation method across simulations in Figure 5.2. In the left panel, we report the proportion of simulations where each covariate was included in the model chosen using the EPE. In the right panel, for each covariate (denoted by color) we report the difference in the proportion of times it was selected for the optimal model using each prediction model versus the proportion of times using the EPE. The EPE estimates are on the x-axis and the difference in proportion is on the y-axis. If a given covariate was included in the model chosen by the EPE and the estimate of the EPE at the same rate, the difference in proportion would be zero.

Next, we compared the estimates of the EPE across 250 simulations. We did this my fixing the order in which covariates were added to the model. We ordered covariates based on how often they were included in the chosen model using the EPE in the first simulation (left panel of Figure 5.2). The covariate most frequently included in the model chosen by EPE was added first and the one least frequently included was added last. We computed the EPE and the estimated EPE using the eight methods for each model and compared the mean estimated EPE for the 8 methods and the mean EPE across simulations in Figure 5.3. We also compared the uncertainty in the EPE estimates by presenting the box plots of the

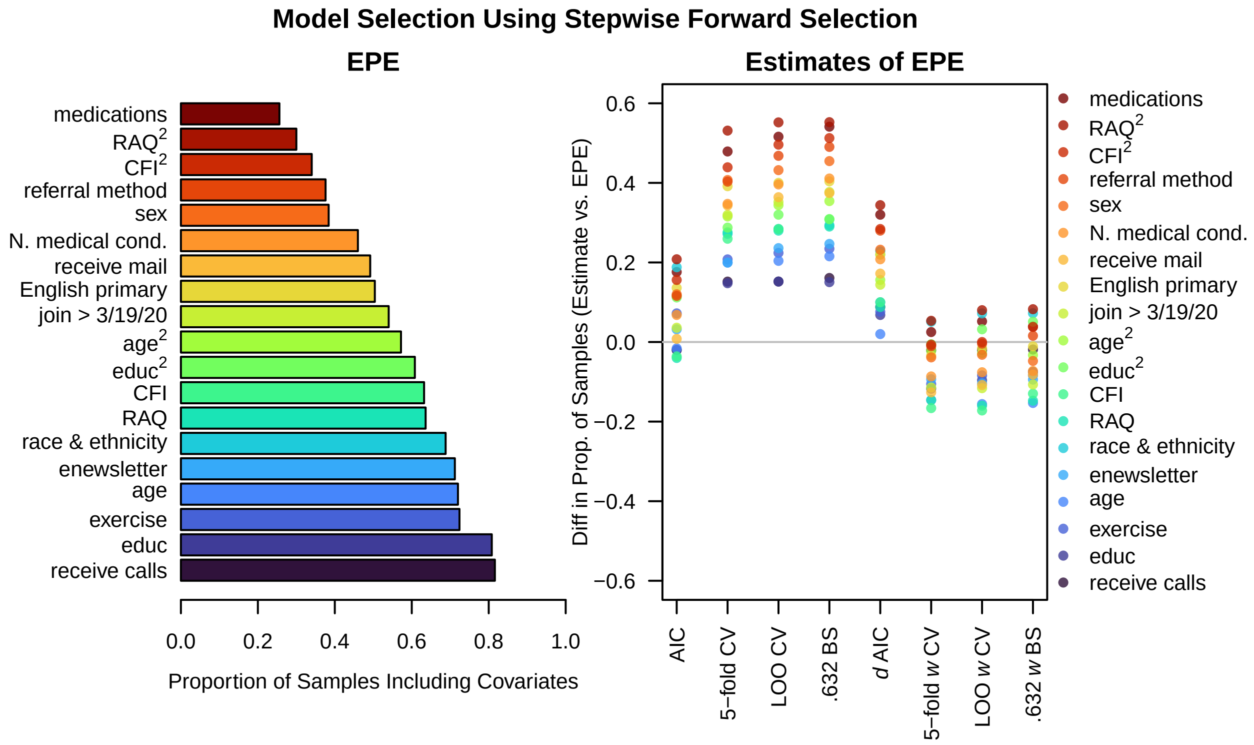


Figure 5.2: The left panel shows the proportion of the 250 simulations each covariate is in the selected model using stepwise forward selection with the expected prediction error (EPE) under log-likelihood loss computed on the test set drawn under a simple random sampling scheme (SRSS). The right panel shows the difference in the proportion of simulations each covariate is in the model selected with each estimate of the EPE versus the one selected with EPE. The 8 methods used to estimate the EPE are AIC, 5-fold cross-validation (5-fold CV), leave-one-out cross-validation (LOO CV), .632 bootstrap (.632 BS), the design-based AIC (dAIC), weighted 5-fold cross-validation (5-fold wCV), weighted leave-one-out cross-validation (LOO wCV), and weighted .632 bootstrap (.632 wBS).

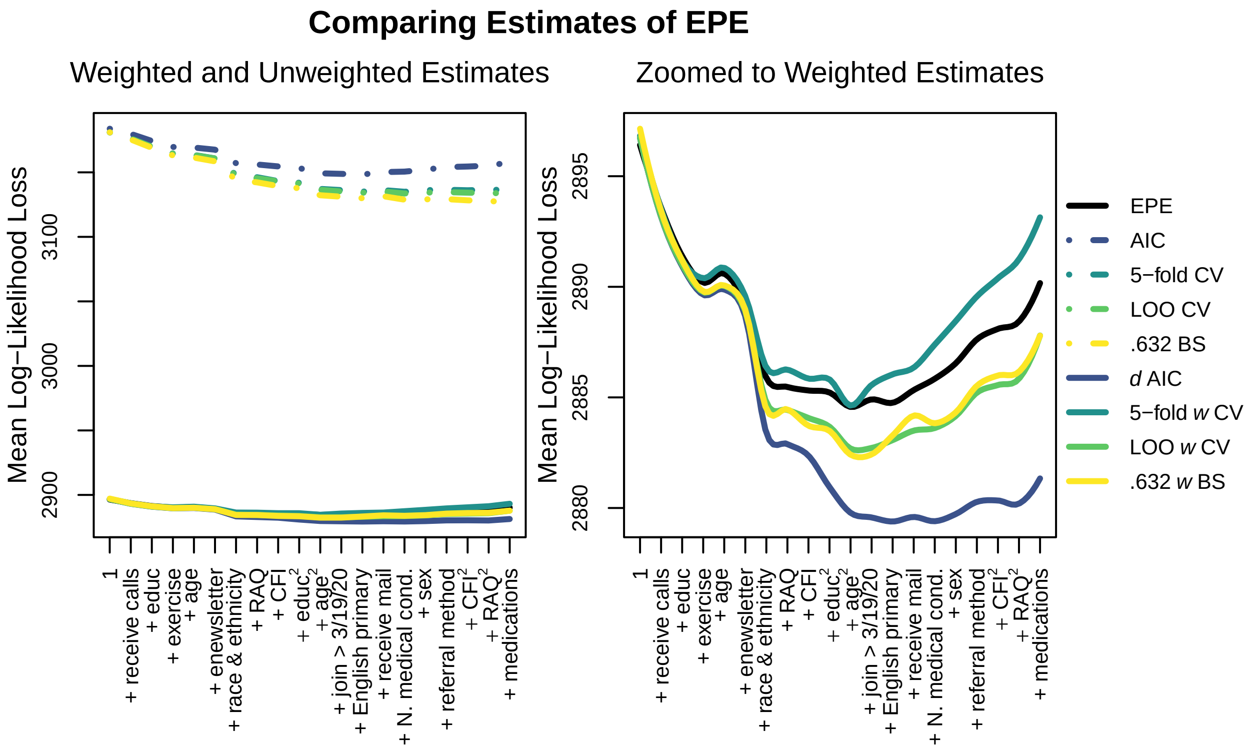


Figure 5.3: The mean expected prediction error [EPE] (black), unweighted estimates of the EPE (dashed colors), and weighted estimates of the EPE (solid colors) are shown for 20 nested models. The EPE is computed on a test set drawn with a simple random sampling scheme and the estimates are computed on a training set drawn with a biased sampling scheme. The results are based on 250 simulations. The right panel shows the same data as the left panel, but it is zoomed into the weighted estimates to show more detail.

distribution of the difference between the estimates and the mean EPE for each method and each model in Figure 5.4.

Third, we assessed the impact of the sample size on each methods performance. In Section 5.4 we discussed how AIC and dAIC estimates should converge to the LOO CV estimates as the sample size increases. We fixed the model to include all possible predictors and varied the sample size n of the training set (sampled under a BSS) and the test set (sampled with a SRSS). We conducted 250 simulations and reported the mean estimated EPE for all methods for n ranging from 500 to $10,000$ in Figure 5.5.

Lastly, we investigated the consequences of estimating sampling weights for a convenience

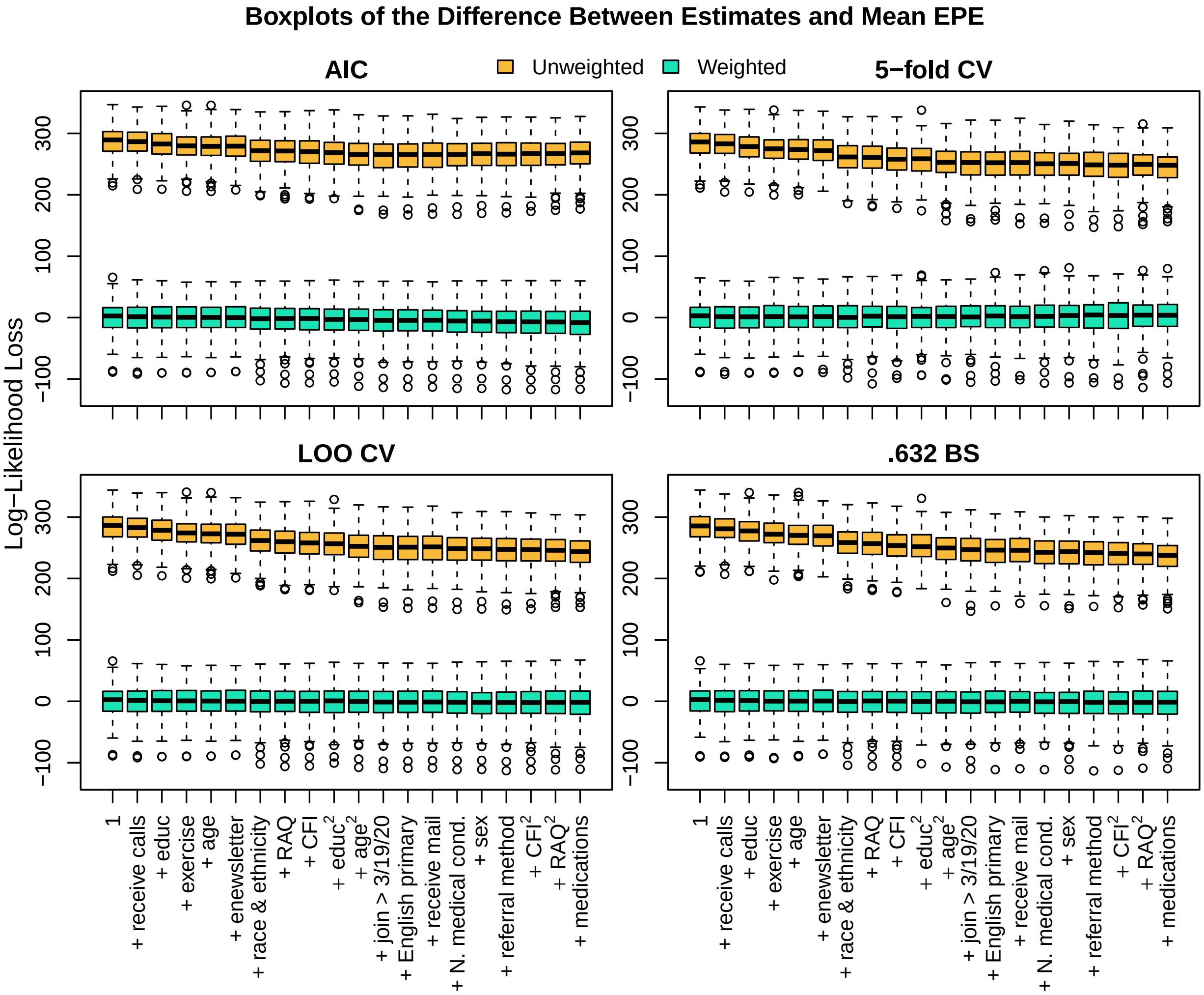


Figure 5.4: The distribution of the difference between the unweighted estimates (orange) and weighted estimates (green) of the EPE versus the mean EPE are shown for 20 nested models. The EPE is computed on a test set drawn with a simple random sampling scheme and the estimates are computed on a training set drawn with a biased sampling scheme. The results are based on 250 simulations.

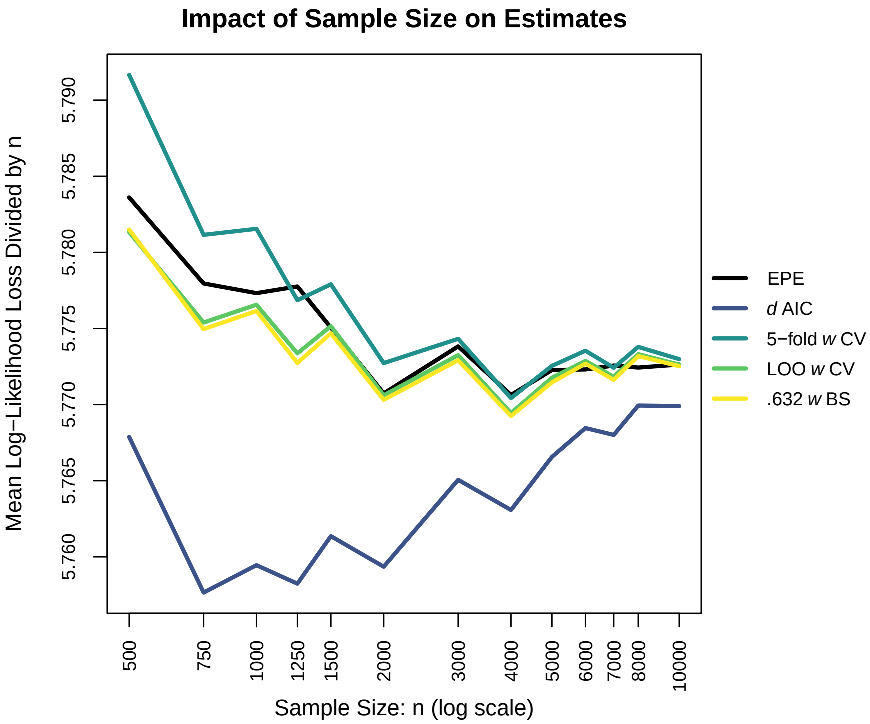


Figure 5.5: The mean expected prediction error [EPE] (black) and weighted estimates of the EPE (colors) are shown for varying sizes (n) of the training and test sets. The EPE is computed on a test set drawn with a simple random sampling scheme and the estimates are computed on a training set drawn with a biased sampling scheme. The results are based on 250 simulations.

samples instead of where the sampling probabilities were not prespecified. We drew an additional sample drawn with a SRSS of size 500 from the population to act as a representative sample for estimating sampling weights. We estimated the probability of inclusion in the training sample versus the representative sample, π_i, using logistic regression with age, education, medications, sex, race and ethnicity, N. medical conditions, exercise, and ADI as covariates. We included quadratic terms in the scope and used stepwise forward selection with AIC to select a logistic regression model. We used the estimated inclusion probabilities to compute sampling weights $\widehat{w}_i \propto (1-\pi_i)\pi_i^{-1}$ as discussed in our previous work [15]. Recall that ADI is the only covariate related to the sampling probability so we have misspecified our sampling probability model by including additional covariates. Thus, we should obtain a reasonable, but not perfect, estimate of the sampling weights. We followed the same order of adding variables as in Figure 5.3 and fit weighted linear regression models and estimated the prediction error for each method. We conducted 250 simulations and reported the mean estimated EPE for each method and each model in Figure 5.6.

5.5.2 Simulation results

In the left panel of Figure 5.2 we see that when the EPE is computed on the test set, the indicator of whether participants agreed to receiving calls from the C2C was in the selected model in more than 80% of the simulations and was the most frequently selected covariate. The indicator of prescription drug use was the least frequently selected covariate. These results are reasonable because the receive calls indicator has the largest coefficient in the true model (Equation 5.8). In the right panel, we are able to assess the frequency with which each covariate is selected into the chosen model when the EPE estimates are used versus the true EPE. The unweighted estimates are the four estimates on the left side of the plot. We see that the models selected using unweighted estimates of the EPE included most covariates more frequently than the models selected with the true EPE. In this

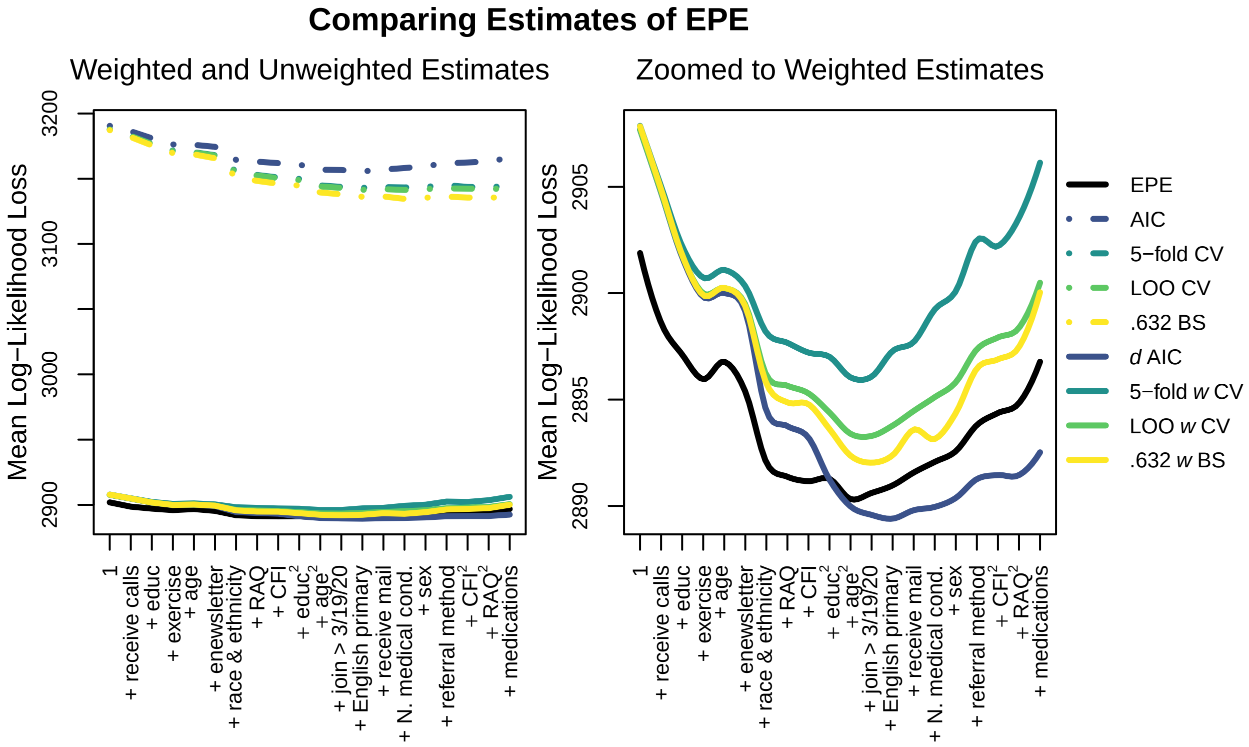


Figure 5.6: The mean expected prediction error [EPE] (black), unweighted estimates of the EPE (dashed colors), and weighted estimates of the EPE (solid colors) are shown for 20 nested models when sampling weights are estimated with a reference sample. Details on the sampling weight estimation is in Section 5.5.1. The EPE is computed on a test set drawn with a simple random sampling scheme and the estimates are computed on a training set drawn with a biased sampling scheme. The results are based on 250 simulations. The right panel shows the same data as the left panel, but it is zoomed into the weighted estimates to show more detail.

context, ignoring the sampling scheme led to overfit models. Models selected using weighted estimates of the EPE tend to include covariates at the same rate as the model selected with the true EPE. Additionally, there is more variability in the difference in proportion for the unweighted estimates than the weighted estimates. In 33 out of the 250 simulations, the bootstrap was unable to estimate the prediction error for one or more models due to sparsity in the bootstrap samples and in 1 simulation the 5-fold CV was unable to estimate the EPE for one model. The unweighted and weighted BS and CV estimates were fit with the same bootstrap samples or folds, respectively so the the rate was the same for the unweighted and weighted estimates. The sparsity was generally caused by the race and ethnicity covariate because only 1.2% of C2C participants report being NH Black.

The results in Figure 5.2 for the dAIC and AIC are flipped from the other estimates. The AIC tends to pick models more similar to those selected by the EPE compared to those selected with the dAIC, but this fits with the results shown in Figure 5.3. The left panel in Figure 5.3 shows the mean EPE estimate as covariates are added to the model for the unweighted (dashed lines) and weighted (solid lines) estimates. The true EPE is shown in black and the estimates are shown in other colors. To highlight differences in the weighted estimators, the right panel zooms into the weighted estimates and the true EPE. Overall, we see that the weighted estimators provide a better estimate of the true EPE. When we compare the weighted estimates to each other in the right panel, we see that the resampling-based estimates are closer to the true EPE than the dAIC.

The prediction error curve for the true EPE is minimized when 10 covariates are added to the model (up to and including age^2). The LOO wCV, .632 wBS, and 5-fold wCV curves have the same minimum, but the dAIC estimate of the loss function is minimized when 14 covariates are added (up to the number of medical conditions). The LOO CV, .632 BS, and 5-fold CV prediction error curves are minimized when 18, 19, and 17 covariates are included, respectively. The AIC curve is minimized when 12 covariates are added. Although the dAIC is a better estimate of the true EPE than the AIC, the AIC is minimized at a model that is more similar to the model that minimizes the true EPE.

Figure 5.4 shows the distribution of the difference between the EPE estimates and the EPE across simulations. Each of the four panels shows a different estimation method: AIC/dAIC, LOO CV/LOO wCV, .632 BS/.632 wBS, and 5-fold CV/5-fold wCV. The unweighted estimates are shown in orange and the weighted estimates in green. The weighted estimates are approximately unbiased and centered at zero but the unweighted estimates are biased upwards. The weighted estimates have more uncertainty than the unweighted estimates. For this simulation, the bootstrap was unable to estimate the EPE due to sparsity for at least one model in 41 out of 250 simulations and 5-fold CV was unable to estimate the EPE in 2 simulations.

In Figure 5.5 we can assess the impact of the sample size. The sample size is on the x-axis, the mean negative log-likelihood loss divided by the sample size n is on the y-axis, and the different estimates are denoted by varying colors. The mean estimates are converging towards the true EPE as the sample size increases. In this simulation scenario, the bootstrap was unable to estimate the EPE in 7 out of 250 simulations and 5-fold CV was unable to estimate the EPE in 1 simulation and only when the sample size was 500. In this study, the sparsity was only a problem with the smallest sample size so it should be less problematic for larger sample sizes.

Figure 5.6 shows the mean estimated EPE by estimation method when the sampling weights are estimated. We see very similar results to when the weights were known. The weighted estimates are less biased for the EPE than the unweighted estimates.

5.6 Application

5.6.1 Application methods

We used the same C2C data set described above in Section 5.5, but we did not simulate a response variable. C2C participants were asked if they were willing to be contacted about studies that involve approved medications, investigational medications, diet and lifestyle interventions, blood draws, cognitive tests, magnetic resonance imaging, Positron Emission Tomography (PET scans), lumbar punctures, and autopsies. We created a willingness score that counted the number of study types participants agreed to be contacted about. The score ranged from 0 to 9. We did subject wise imputation for those who were missing less than 30% of the questions about willingness (1 or 2 questions). We replaced the missing values with the participant's average score on the other questions. We treated the total willingness score as missing for those who did not answer 3 or more willingness questions. There were 10 participants missing 3 or more questions.

We treated the complete C2C data set as our population of interest and drew a test set using a SRSS of size 500 to estimate the true EPE. We then drew a training set with replacement of size 500 from the remaining observations with a BSS using sampling probabilities for each ADI score inversely proportional to the proportion of that score in C2C. We used the training sample to fit weighted linear regression models to predict willingness. We used the same predictors as were used in the simulation study. We performed stepwise forward selection to add covariates sequentially using the EPE computed on the test set and used stepwise forward selection to add all covariates. We determined the next covariate to add choosing the one with the lowest estimate prediction error. We then estimated the prediction error with the unweighted (AIC, LOO CV, .632 BS, 5-fold CV) and the weighted (dAIC, LOO wCV, .632 wBS, and 5-fold wCV) methods for all of the models selected using stepwise forward selection.

We summarized continuous covariates with means and standard deviations and the categorical covariates with proportions for the full C2C data set that we used as the finite population. We similarly summarized the covariates in the two samples drawn with a SRSS and BSS. Additionally, we presented the distribution of the willingness scores using a histogram for the samples drawn with a SRSS and BSSS.

5.6.2 Application results

There were 3,773 observations in the full data set and after removing missing data there were 2,822 observations. 725 participants were only missing an ADI score, 27 were only missing RAQ, and 11 were missing both. There were more missing values for ADI because it could not be computed without a valid address or for PO boxes.

The summarized covariates for the full C2C data set and the samples drawn with a SRSS and BSS are presented in Figure 5.1. Individuals in the sample drawn with a BSS tend to be slight older, have fewer years of education, and have a higher CFI on average, compared to the C2C data set and the sample drawn with a SRSS. Additionally, individuals in the sample drawn with a BSS are less likely to identify as NH White and are more likely to identify as female than those in the C2C data set and the sample drawn with a SRSS.

A histogram of the willingness scores stratified by sampling scheme is shown in Figure 5.7. Higher willingness scores correspond to agreeing to be contacted about more types of studies. The distribution of willingness scores across samples are similar, but the individuals in the sample drawn with a BSS tend to have slightly higher scores.

The true EPE and the estimates for the four unweighted and four weighted methods are shown in Figure 5.8. The weighted estimates are closer to the true EPE than the unweighted estimates. The weighted estimates seem to be more variable. All estimates of the EPE reach

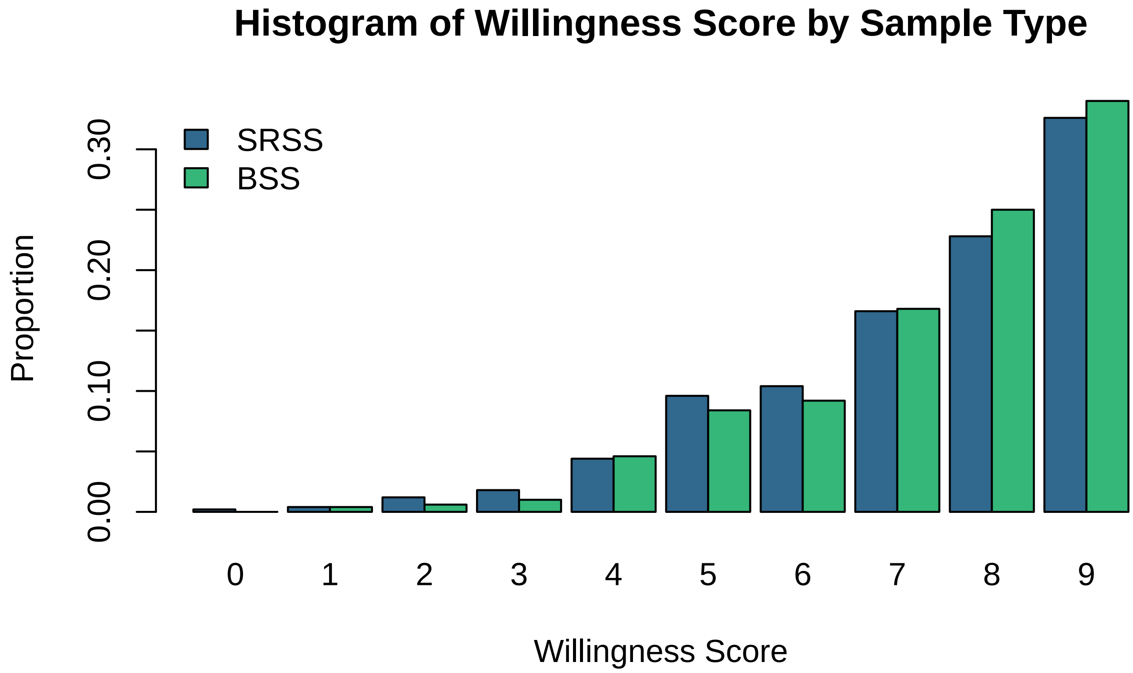


Figure 5.7: Histogram of willingness scores stratified by sampling scheme. The sample drawn with a simple random sampling scheme (SRSS) is shown in blue and the sample drawn with a biased sampling scheme (BSS) is shown in green.

	Finite Population	Samples	
	C2C (N = 2,822)	SRSS (n = 500)	BSS (n = 500)
Age	58.4 (15.9)	57.6 (16.1)	57.1 (16.6)
Education (years)	16.4 (2.7)	16.3 (2.6)	16.0 (3.0)
CFI	3.0 (3.0)	3.1 (3.0)	3.3 (3.3)
RAQ	28.6 (4.4)	28.5 (4.6)	28.7 (4.1)
Race & Ethnicity: Hispanic	0.10	0.10	0.13
Race & Ethnicity: NH Asian	0.06	0.07	0.07
Race & Ethnicity: NH Black	0.01	0.01	0.02
Race & Ethnicity: NH White	0.73	0.70	0.67
Race & Ethnicity: Other	0.09	0.11	0.11
Female	0.64	0.61	0.69
Enrollment after 3/19/20	0.06	0.08	0.04
English Primary	0.98	0.96	0.97
Exercise	0.85	0.85	0.82
Receive Calls	0.67	0.68	0.69
Receive Mail	0.84	0.83	0.84
Enewsletter	0.71	0.73	0.73
N. Medical Conditions: 0	0.11	0.12	0.12
N. Medical Conditions: 1	0.64	0.64	0.61
N. Medical Conditions: 2+	0.25	0.24	0.26
Referral Method: Email	0.54	0.54	0.55
Referral Method: Community Talk	0.10	0.08	0.10
Referral Method: Other	0.36	0.38	0.35

Table 5.1: Continuous covarates are summarized using means (standard deviations) and categorical variables are summarized using proportions for the full C2C data set and samples drawn with a simple random sampling scheme (SRSS) and a biased sampling scheme (BSS).

their minimum at a smaller model than the model with the minimum EPE.

5.7 Discussion

In this chapter, we have discussed several examples of weighted analytic and resampling methods for estimating the expected prediction error in a target population for samples drawn with a biased sampling scheme. The results from our simulation study show that it

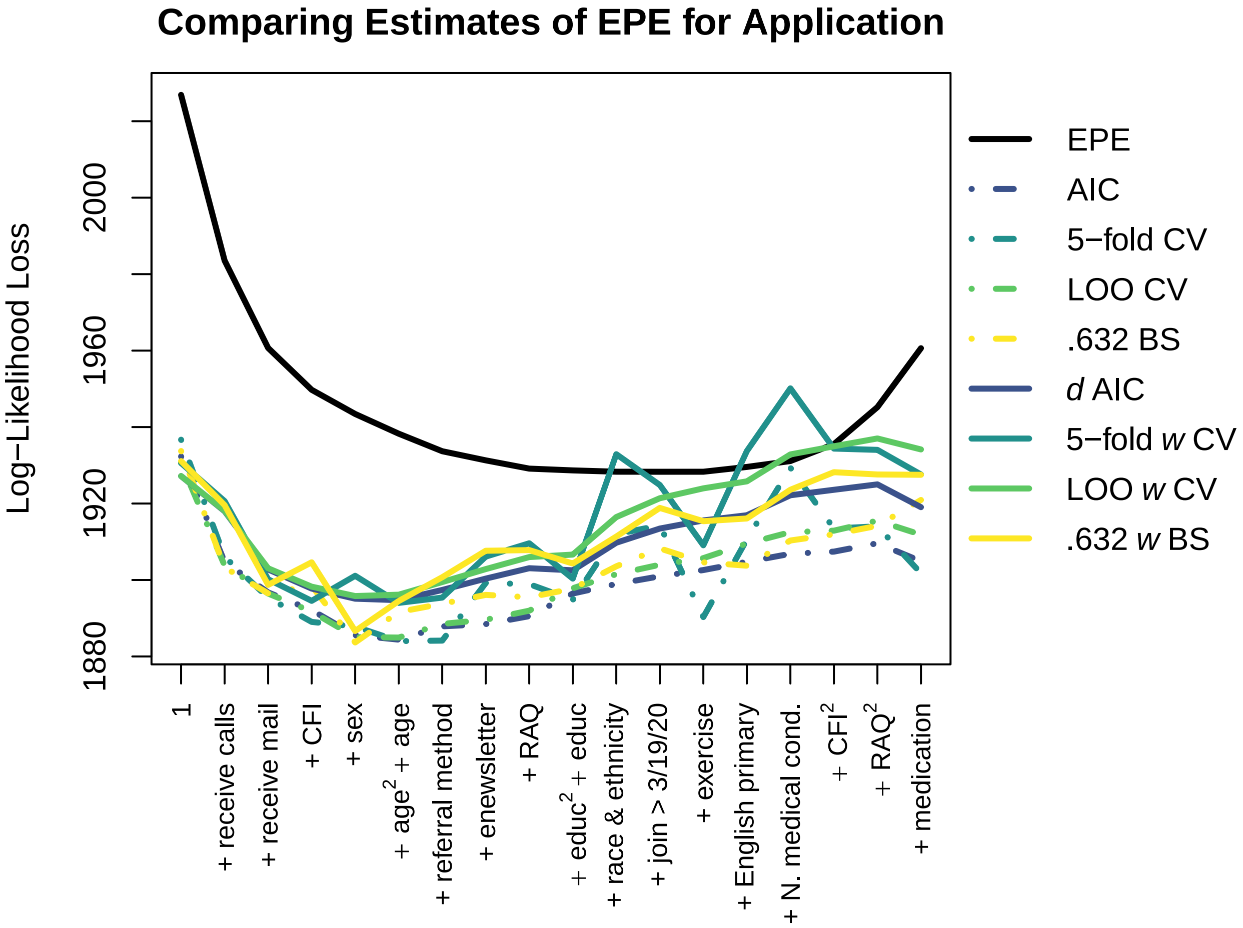


Figure 5.8: The expected prediction error [EPE] (black), unweighted estimates of the EPE (dashed colors), and weighted estimates of the EPE (solid colors) are shown for 20 nested models for predicting research willingness among C2C participants. The EPE is computed on a hold out test set drawn with a simple random sampling scheme and the estimates are computed on a training set drawn with replacement using a biased sampling scheme. Results are shown for one training and test set drawn from C2C.

is important to account for the sampling scheme by including sampling weights in estimates of the prediction error (5.3). Weighted estimates of the EPE in the simulation study were more variable (Figure 5.4), but this is common with weighted estimates. This occurs because subpopulations with small sampling probabilities and large weights have a large influence on estimates [15, 71]. It is better to have a less precise unbiased estimate than a precise biased estimate. Weighted estimates, excluding the dAIC, performed better for model selection than unweighted estimates (Figure 5.2). Although weighted estimates of the EPE are more variable, there is less variability in which model they select.

Additionally, in the simulation study we showed that weighted resampling estimates are less biased for the EPE than weighted analytic estimates. These results were expected because resampling methods directly estimate the EPE and do not assume the distribution of the covariates is fixed [54]. Analytic estimates were consistent for the same quantity, however, and performed better as the sample size increased. This is consistent with the theory derived by Lumley and Scott [76].

In practice, based on previously developed theoretical results and our simulation study we recommend using weighted resampling methods for estimating the prediction error for a BSS if it is computationally feasible. Analytic methods, however, still perform very well. Although the dAIC was more biased for the EPE than the resampling methods it was still a reasonable estimate (Figure 5.3). It did not outperform the unweighted AIC for model selection, but that may just be an artifact of our simulation scenario because it was a better estimate of the EPE. In general, estimating the EPE well should translate to doing well at model selection. Weighted analytic estimates of the EPE, such as the dAIC are a great option for larger samples because they perform better as the sample size increases which is when the computational burden increases for resampling methods. If researchers are interested in using a weighted analytic method with a loss function besides log-likelihood loss, they can use the Horvitz-Thompson-Efron estimator [57]. When deciding between resampling methods, 5-fold cross-validation is less likely to face sparsity than the bootstrap which often has 100 bootstrap samples drawn with replacement. LOO CV is a good option when there is an analytic form, as in our scenario, but it would be more computationally expensive than the 5-fold CV otherwise. Thus, we suggest using the weighted 5-fold cross-validation to estimate the EPE if it is computationally feasible.

In our application of assessing prediction error for prediction models of research willingness using C2C data, we observed that weighted estimates were less biased than unweighted estimates. The differences between these two groups were less stark than the differences in the simulation study, but it is still important to account for the sampling scheme. In future research projects, we can not rule out that sampling weights could have a large impact on the EPE estimates. Sampling weights will matter if the relationship between the predictors and research willingness is modified by a variable related to ADI. Socioeconomic disadvantage is thought of as a social determinant of health [18] and C2C researchers want to gain more information on recruitment strategies for individuals traditionally represented in research [87]. When the C2C recruitment strategy is updated in the future to oversample neighborhoods with higher degrees of socioeconomic disadvantage, sampling weights need to be included when fitting assessing predictive models trained with C2C data.

Our results are limited by how accurate the sampling weights are. This limitation is obvious for convenience samples where weights are estimated. Estimated sampling weights are subject to an analogue of confounding if covariates related to the sampling scheme are not measured. For more details on the assumptions see our previous work [15]. For example, even without oversampling neighborhoods with more socioeconomic disadvantage, the C2C tends to underrepresent groups historically underrepresented in medical research [15, 87]. We can estimate sampling weights for the C2C to generalize to the US population, but we may not have measured all covariates related to the sampling probability. This is also a limitation in survey samples with prespecified sampling probabilities. The true sampling probabilities may differ from the prespecified probabilities due to sampling variability or non-response [75]. Accounting for sampling weights that are not perfect, but still contain information about the sampling scheme is better than failing to account for the sampling scheme.

In conclusion, all weighted estimates of the expected prediction error for the target population under a BSS perform well. The difference between unweighted and weighted estimates is larger than the difference between weighted resampling methods and weighted analytic methods. Resampling methods perform better than analytic methods because they directly estimate the EPE, but they are more computationally expensive. We used weighted linear regression models as an example in this chapter, but these relationships should hold for more flexible weighted prediction models such as random forest [19] or generalized linear models [85]. We recommend using resampling methods in smaller samples, but we recommend analytic methods in larger samples when they get more computationally expensive and the difference between analytic and resampling methods decreases. Among weighted resampling methods, we suggest using 5-fold cross-validation because it requires less resampling which saves computational cost and has fewer issues with sparsity. Accounting for the sampling scheme by using a weighted estimate is more important than the decision between analytic and resampling methods.

Chapter 6

Discussion and Future Research Directions

In this dissertation, we proposed and assessed methods for estimating associations and predictions that generalize to a set target population when using a biased sample. We accomplished this by estimating sampling weights and quantifying uncertainty in associations for biased samples in biomedical applications using NHANES as a representative sample of the United States. Additionally, we assessed the utility of incorporating estimated sampling weights in propensity score adjusted estimates of causal effects and quantified uncertainty in the estimates. Finally, we compared analytic and resampling estimates of prediction error for biased samples.

In Chapter 3, we discussed previous work on estimating sampling weights for biased samples and the necessity of obtaining a representative sample. We proposed using NHANES as practical solution for this requirement for biomedical applications. We derived an analytic variance estimate to quantify uncertainty in coefficient estimates in weighted generalized linear models for biased samples. We developed an R package with the goal of reducing the amount of work necessary to account for the sampling scheme when fitting GLMS. Our `estweight` package contains functions to (1) estimate sampling weights using four predictive models (logistic regression, covariate balancing propensity score, entropy balancing, and random forest), (2) estimate coefficient estimates for weighted GLMs for biased samples, and (3) provide variance estimates.

In Chapter 4, we assessed when and how estimated sampling weights should be included when estimating propensity scores and propensity-adjusted causal estimates for biased samples– particularly when the propensity score model was misspecified. We found that the magnitude of treatment effect heterogeneity (if not correctly modeled) impacted the degree of sampling bias. As the heterogeneity increased, so did the bias. Previous papers agreed that sampling weights needed to be included in the outcome model, but disagreed about the necessity of including them in the propensity score model. We observed that failing to include sampling weights in the propensity score model led to increased volatility of causal estimates. We also assessed the impact of how well sampling weights were estimated and derived a variance estimate for the estimated causal effect that accounted for uncertainty from estimating the sampling weights, propensity scores, and the causal effect. We extended our `estweight` package for R to include functions to estimate sampling weights, propensity scores, propensity-adjusted causal effects and corresponding variance estimates.

Lastly, in Chapter 5, we compared different methods for estimating the prediction error for a target population with a biased sample. We found that prediction estimates that included sampling weights were less biased than unweighted estimates. Within weighted estimates, resampling methods outperformed analytic estimates. The analytic estimates, however, are less computationally expensive and converge to the resampling methods as the sample size increases. We recommend using resampling methods when computationally feasible, and more specifically we recommend weighted 5-fold cross-validation because it requires less computation than weighted leave-one-out cross-validation (except for cases with an analytic form) and has less difficulty from sparsity than the bootstrap. We recommend using an analytic estimate like the design-based estimate or Horvitz-Thompson-Efron estimator for larger sample sizes.

There is more work to be done in addressing sampling bias. In this next section we discuss future research areas.

6.1 Future Work

The work presented in this dissertation can be extended in several different directions. First, there is need of sensitivity analyses to assess the impact of missing variables in the sampling weights estimation model. Second, the design-based AIC assumes the sampling weights are prespecified but could be expanded to convenience samples where the weights are estimated with an auxiliary data set.

Similar to the unconfounded assumption for estimating causal effects, using estimated sampling weights to address sampling bias is based on the assumption that covariates related to the sampling probability are measured (see discussion in Section 3.2.1). When we submitted the materical in Chapter 3 for publication, one of the reviewers asked for a sensitivity analysis of the impact of missing variables in the sampling weight estimation model for our applied data example. We decided to save this project for future work because to our knowledge there were no existing method to handle this. In Chapter 4, we used a simulation study to assess the results of failing to fully capture all variables related to the sampling scheme (see Figure 4.4). We considered a scenario where one variable determined the sampling probability, X_1, but instead of measuring it we only observed a proxy, K. We found that the weaker the correlation between X_1 and K, the more bias there was in the estimated treatment effect. This analysis was done with simulated data, however, and the reviewer suggested a sensitivity analysis for real data. It would be useful to develop something similar to Cornfield at al. (1959) [29] to assess the impact on magnitude of estimated coeffients for the scientific outcome model from a missing confounder in the sampling weight estimation model.

Next, the dAIC could be extended for convenience samples with estimated sampling weights. The dAIC is a function of a consistent estimate of the covariance of the estimating function. The current covariance estimate that is implemented assumes sampling weights are known, but this could be replaced with an estimate that assumes weights are estimated as proposed in Chapter 3.

Bibliography

[1] Jason Abrevaya, Yu-Chin Hsu, and Robert P. Lieli. Estimating Conditional Average Treatment Effects. *Journal of Business & Economic Statistics*, 33(4):485–505, October 2015. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/07350015.2014.975555.

[2] Benjamin Ackerman, Catherine R. Lesko, Juned Siddique, Ryoko Susukida, and Elizabeth A. Stuart. Generalizing randomized trial findings to a target population using complex survey population data. *Statistics in Medicine*, 40(5):1101–1120, 2021. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/sim.8822.

[3] Alan Agresti. *Foundations of Linear and Generalized Linear Models | Wiley*. John Wiley & Sons, 2015.

[4] Hirotugu Akaike. Information theory and an extension of the maximum likelihood principle. *Second International Symposium on Information Theory*, pages 267–281, 1973.

[5] Hirotugu Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, December 1974. Conference Name: IEEE Transactions on Automatic Control.

[6] David M. Allen. The Relationship between Variable Selection and Data Agumentation and a Method for Prediction. *Technometrics*, 16(1):125–127, 1974. Publisher: [Taylor & Francis, Ltd., American Statistical Association, American Society for Quality].

[7] Rebecca E. Amariglio, Michael C. Donohue, Gad A. Marshall, Dorene M. Rentz, David P. Salmon, Steven H. Ferris, Stella Karantzoulis, Paul S. Aisen, and Reisa A. Sperling. Tracking early decline in cognitive function in older individuals at risk for Alzheimer's disease dementia: the Alzheimer's Disease Cooperative Study Cognitive Function Instrument. *JAMA neurology*, 72(4):446–454, April 2015.

[8] Peter C. Austin. An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies. *Multivariate Behavioral Research*, 46(3):399–424, May 2011.

[9] Peter C Austin, Nathaniel Jembere, and Maria Chiu. Propensity score matching and complex surveys. *Statistical Methods in Medical Research*, 27(4):1240–1257, April 2018.

[10] Peter C Austin and Elizabeth A Stuart. Estimating the effect of treatment on binary outcomes using full matching on the propensity score. *Statistical Methods in Medical Research*, 26(6):2505–2525, 2015. Publisher: SAGE Publications Ltd STM.

[11] L. Charles Bailey, David E. Milov, Kelly Kelleher, Michael G. Kahn, Mark Del Beccaro, Feliciano Yu, Thomas Richards, and Christopher B. Forrest. Multi-Institutional Sharing of Electronic Health Record Data to Assess Childhood Obesity. *PloS One*, 8(6):e66192, 2013.

[12] Luta Luse Basambombo, Pierre-Hugues Carmichael, Sharlène Côté, and Danielle Laurin. Use of Vitamin E and C Supplements for the Prevention of Cognitive Decline. *The Annals of Pharmacotherapy*, 51(2):118–124, February 2017.

[13] Duane L. Beekly, Erin M. Ramos, William W. Lee, Woodrow D. Deitrich, Mary E. Jacka, Joylee Wu, Janene L. Hubbard, Thomas D. Koepsell, John C. Morris, Walter A. Kukull, and The NIA Alzheimer's Disease Centers. The National Alzheimer's Coordinating Center (NACC) Database: The Uniform Data Set. *Alzheimer Disease & Associated Disorders*, 21(3):249–258, July 2007.

[14] Dennis S. Bernstein. Basic Matrix Properties. In *Matrix Mathematics: Theory, Facts, and Formulas*, page 159. Princeton University Press, Princeton, New Jersey, 2 edition, 2009.

[15] Olivia M. Bernstein, Brian G. Vegetabile, Christian R. Salazar, Joshua D. Grill, and Daniel L. Gillen. Adjustment for biased sampling using NHANES derived propensity weights. *Health Services and Outcomes Research Methodology*, July 2022.

[16] David A. Binder. On the Variances of Asymptotically Normal Estimators from Complex Surveys. *International Statistical Review / Revue Internationale de Statistique*, 51(3):279–292, 1983. Publisher: [Wiley, International Statistical Institute (ISI)].

[17] Y. M. M. Bishop, S. E. Fienberg, and P. W. Holland. *Discrete Multivariate Analysis: Theory and Practice*. MIT Press, Cambridge, 1975.

[18] Paula Braveman and Laura Gottlieb. The Social Determinants of Health: It's Time to Consider the Causes of the Causes. *Public Health Reports*, 129(1_suppl2):19–31, 2014. Publisher: SAGE Publications Inc.

[19] Leo Breiman. Random Forests. *Machine Learning*, 45(1):5–32, October 2001.

[20] Leo Breiman, Jerome Friedman, Charles J. Stone, and R. A. Olshen. *Classification and Regression Trees*. Chapman and Hall/CRC, Boca Raton, 1st edition edition, January 1984.

[21] Leo Breiman and Philip Spector. Submodel Selection and Evaluation in Regression. The X-Random Case. *International Statistical Review / Revue Internationale de Statistique*, 60(3):291–319, 1992. Publisher: [Wiley, International Statistical Institute (ISI)].

[22] Declan Browne, Bernadette McGuinness, Jayne V Woodside, and Gareth J McKay. Vitamin E and Alzheimer's disease: what do we know so far? *Clinical Interventions in Aging*, 14:1303–1317, July 2019.

[23] George Casella and Roger Lee Berger. *Statistical Inference*. Brooks/Cole Cengage Learning, Belmont, CA, second edition, 2002.

[24] G.C. Cawley. Leave-One-Out Cross-Validation Based Model Selection Criteria for Weighted LS-SVMs. *The 2006 IEEE International Joint Conference on Neural Network Proceedings*, pages 1661–1668, July 2006.

[25] Centers for Disease Control and Prevention (CDC). National Health and Nutrition Examination Survey Data (2013-2016), 2013.

[26] Yilin Chen, Pengfei Li, and Changbao Wu. Doubly Robust Inference With Nonprobability Survey Samples. *Journal of the American Statistical Association*, 115(532):2011–2021, October 2020. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/01621459.2019.1677241.

[27] Yang Cheng, Eric Slud, and Carma Hogue. Variance Estimation for Decision-Based Estimators with Application to the Annual Survey of Public Employment and Payroll. *Governments Division Report Series*, Research Report #2010-3, 2010. U.S. Census Bureau.

[28] Gerda Claeskens and Nils Lid Hjort. *Model Selection and Model Averaging*. Cambridge University Press, 2008. Publication Title: Cambridge Books.

[29] J. Cornfield, W. Haenszel, E. C. Hammond, A. M. Lilienfeld, M. B. Shimkin, and E. L. Wynder. Smoking and lung cancer: recent evidence and a discussion of some questions. *Journal of the National Cancer Institute*, 22(1):173–203, January 1959.

[30] Rajeev H. Dehejia and Sadek Wahba. Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs. *Journal of the American Statistical Association*, 94(448):1053–1062, December 1999. Publisher: Taylor & Francis _eprint: https://www.tandfonline.com/doi/pdf/10.1080/01621459.1999.10473858.

[31] Christiana Drake. Effects of Misspecification of the Propensity Score on Estimators of Treatment Effect. *Biometrics*, 49(4):1231–1236, 1993. Publisher: [Wiley, International Biometric Society].

[32] Eva H. Dugoff, Megan Schuler, and Elizabeth A. Stuart. Generalizing observational study results: applying propensity score methods to complex surveys. *Health Services Research*, 49(1):284–303, February 2014.

[33] Maurice W. Dysken, Mary Sano, Sanjay Asthana, Julia E. Vertrees, Muralidhar Pallaki, Maria Llorente, Susan Love, Gerard D. Schellenberg, J. Riley McCarten, Julie Malphurs, Susana Prieto, Peijun Chen, David J. Loreck, George Trapp, Rajbir S. Bakshi, Jacobo E. Mintzer, Judith L. Heidebrink, Ana Vidal-Cardona, Lillian M. Arroyo, Angel R. Cruz, Sally Zachariah, Neil W. Kowall, Mohit P. Chopra, Suzanne

Craft, Stephen Thielke, Carolyn L. Turvey, Catherine Woodman, Kimberly A. Monnell, Kimberly Gordon, Julie Tomaska, Yoav Segal, Peter N. Peduzzi, and Peter D. Guarino. Effect of Vitamin E and Memantine on Functional Decline in Alzheimer Disease: The TEAM-AD VA Cooperative Randomized Trial. *JAMA*, 311(1):33–44, January 2014.

[34] Bradley Efron. Bootstrap methods: another look at the jackknife. *Annals of Statistics*, 7:1–26, 1979.

[35] Bradley Efron. Estimating the Error Rate of a Prediction Rule: Improvement on Cross-Validation. *Journal of the American Statistical Association*, 78(382):316–331, June 1983. Publisher: Taylor & Francis _eprint: https://www.tandfonline.com/doi/pdf/10.1080/01621459.1983.10477973.

[36] Bradley Efron, Prabir Burman, L. Denby, J. M. Landwehr, C. L. Mallows, Xiaotong Shen, Hsin-Cheng Huang, Jianming Ye, Jimmy Ye, and Chunming Zhang. The Estimation of Prediction Error: Covariance Penalties and Cross-Validation [with Comments, Rejoinder]. *Journal of the American Statistical Association*, 99(467):619–642, 2004.

[37] Bradley Efron and R. J. Tibshirani. *An Introduction to the Bootstrap.* CRC Press, 1993. Google-Books-ID: gLlpIUxRntoC.

[38] Bradley Efron and Robert Tibshirani. Improvements on Cross-Validation: The 632+ Bootstrap Method. *Journal of the American Statistical Association*, 92(438):548–560, June 1997. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/01621459.1997.10474007.

[39] Michael R. Elliott, Trivellore E. Raghunathan, and Nathaniel Schenker. Combining Estimates from Multiple Surveys. In *Wiley StatsRef: Statistics Reference Online*, pages 1–10. American Cancer Society, 2018.

[40] Michael R. Elliott and Richard W. Valliant. Inference for Nonprobability Samples. *Statistical Science*, 32:249–264, 2017.

[41] David A Freedman. On The So-Called "Huber Sandwich Estimator" and "Robust Standard Errors". *The American Statistician*, 60(4):299–302, November 2006.

[42] Wayne. A. Fuller. *Sampling Statistics.* John Wiley & Sons, Hoboken, New Jersey, 2009.

[43] Luke M. Funk, Ying Shan, Corrine I. Voils, John Kloke, and Lawrence P. Hanrahan. Electronic Health Record Data Versus the National Health and Nutrition Examination Survey (NHANES): A Comparison of Overweight and Obesity Rates. *Medical Care*, 55(6):598–605, 2017.

[44] Elizabeth C. Gearan, Kelley Monzella, Alice Ann Gola, and Holly Figueroa. Adolescent Participants in the School Lunch Program Consume More Nutritious Lunches but Their 24-hour Diets are Similar to Nonparticipants. *Journal of Adolescent Health*, January 2021.

[45] Andrew Gelman and Thomas C. Little. Poststratification Into Many Categories Using Hierarchical Logistic Regression. *Survey Methodology*, 23:127–135, 1997.

[46] V. P. Godambe. An Optimum Property of Regular Maximum Likelihood Estimation. *The Annals of Mathematical Statistics*, 31(4):1208–1211, December 1960.

[47] Gene H. Golub, Michael Heath, and Grace Wahba. Generalized Cross-Validation as a Method for Choosing a Good Ridge Parameter. *Technometrics*, 21(2):215–223, May 1979. Publisher: Taylor & Francis _eprint: https://www.tandfonline.com/doi/pdf/10.1080/00401706.1979.10489751.

[48] Shelly L. Gray, Melissa L. Anderson, Paul K. Crane, John C. S. Breitner, Wayne McCormick, James D. Bowen, Linda Teri, and Eric Larson. Antioxidant vitamin supplement use and risk of dementia or Alzheimer's disease in older adults. *Journal of the American Geriatrics Society*, 56(2):291–295, February 2008.

[49] Rebecca E. Greenblatt, Edward J. Zhao, Sarah E. Henrickson, Andrea J. Apter, Rebecca A. Hubbard, and Blanca E. Himes. Factors associated with exacerbations among adults with asthma according to electronic health record data. *Asthma Research and Practice*, 5(1):1, January 2019.

[50] Sander Greenland, James M. Robins, and Judea Pearl. Confounding and Collapsibility in Causal Inference. *Statistical Science*, 14(1):29–46, February 1999. Publisher: Institute of Mathematical Statistics.

[51] Joshua D. Grill, Dan Hoang, Daniel L. Gillen, Chelsea G. Cox, Adrijana Gombosev, Kirsten Klein, Steve O'Leary, Megan Witbracht, and Aimee Pierce. Constructing a Local Potential Participant Registry to Improve Alzheimer's Disease Clinical Research Recruitment. *Journal of Alzheimer's Disease*, 63(3):1055–1063, January 2018.

[52] Jens Hainmueller. Entropy Balancing for Causal Effects: A Multivariate Reweighting Method to Produce Balanced Samples in Observational Studies. *Political Analysis*, 20(1):25–46, 2012.

[53] Lars Peter Hansen. Large Sample Properties of Generalized Method of Moments Estimators. *Econometrica*, 50(4):1029–1054, 1982.

[54] Trevor Hastie, Robert Tibshirani, and Jerome Freedman. Model Assessment and Selection. In *The Elements of Statistical Learning*, Springer Series in Statistics, pages 219–259. 2nd edition, 2009.

[55] Trevor Hastie, Robert Tibshirani, and Ryan Tibshirani. Best Subset, Forward Stepwise or Lasso? Analysis and Recommendations Based on Extensive Comparisons. *Statistical Science*, 35(4):579–592, November 2020. Publisher: Institute of Mathematical Statistics.

[56] Keisuke Hirano and Guido W. Imbens. The Propensity Score with Continuous Treatments. In Andrew Gelman and Xiao-Li Meng, editors, *Applied Bayesian Modeling and*

Causal Inference from Incomplete-Data Perspectives, Wiley Series in Probability and Statistics, pages 73–84. John Wiley & Sons, Ltd, Chichester, UK, 2004.

[57] Andrew Holbrook, Thomas Lumley, and Daniel Gillen. Estimating prediction error for complex samples. *Canadian Journal of Statistics*, 48(2):204–221, 2020. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/cjs.11527.

[58] Paul W. Holland. Statistics and Causal Inference. *Journal of the American Statistical Association*, 81(396):945–960, 1986. Publisher: [American Statistical Association, Taylor & Francis, Ltd.].

[59] D. G. Horvitz and D. J. Thompson. A Generalization of Sampling Without Replacement From a Finite Universe. *Journal of the American Statistical Association*, 47(260):663–685, 1952.

[60] D. Hume. *An Enquiry Concerning Human Understanding.* Open Court Press, LaSalle, 1748.

[61] Kosuke Imai and Marc Ratkovic. Covariate balancing propensity score. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):243–263, 2014.

[62] Guido W. Imbens and Donald B. Rubin. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction.* Cambridge University Press, Cambridge, 2015.

[63] Amy J.H. Kind and William R. Buckingham. Making Neighborhood-Disadvantage Metrics Accessible — The Neighborhood Atlas. *New England Journal of Medicine*, 378(26):2456–2458, June 2018.

[64] Leslie Kish. *Survey Sampling.* John Wiley & Sons, New York, NY, USA, 1965.

[65] Ron Kohavi. A study of cross-validation and bootstrap for accuracy estimation and model selection. *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, 2:1137–1143, August 1995.

[66] Richard J. Kryscio, Erin L. Abner, Allison Caban-Holt, Mark Lovell, Phyllis Goodman, Amy K. Darke, Monica Yee, John Crowley, and Frederick A. Schmitt. Association of Antioxidant Supplement Use and Dementia in the Prevention of Alzheimer's Disease by Vitamin E and Selenium Trial (PREADViSE). *JAMA Neurology*, 74(5):567–573, May 2017.

[67] Brian K. Lee, Justin Lessler, and Elizabeth A. Stuart. Weight Trimming and Propensity Score Weighting. *PLOS ONE*, 6(3):e18174, March 2011. Publisher: Public Library of Science.

[68] David Lenis, Trang Quynh Nguyen, Nianbo Dong, and Elizabeth A. Stuart. It's all about balance: propensity score matching in the context of complex survey data. *Biostatistics (Oxford, England)*, 20(1):147–163, 2019.

[69] Fan Li, Kari Lock Morgan, and Alan M. Zaslavsky. Balancing Covariates via Propensity Score Weighting. *Journal of the American Statistical Association*, 113(521):390–400, January 2018. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/01621459.2016.1260466.

[70] Andy Liaw and Matthew Wiener. Classification and Regression by randomForest. *R News*, 2(3):18–22, 2002.

[71] Roderick J. A. Little and Donald B. Rubin. Complete-Case and Available-Case Analysis, Including Weighting Methods. In *Statistical Analysis with Missing Data*, pages 41–58. John Wiley & Sons, Ltd, 2014. Section: 3 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/9781119013563.ch3.

[72] Ana Lloret, Mari-Carmen Badía, Nancy J. Mora, Federico V. Pallardó, Maria-Dolores Alonso, and Jose Viña. Vitamin E paradox in Alzheimer's disease: it does not prevent loss of cognition and may even be detrimental. *Journal of Alzheimer's disease: JAD*, 17(1):143–149, 2009.

[73] Sharon L. Lohr. *Sampling: Design and Analysis.* Brooks/Cole Cengage Learning, second edition, 2010.

[74] Jose A. Luchsinger, Ming-Xin Tang, Steven Shea, and Richard Mayeux. Antioxidant vitamin intake and risk of Alzheimer disease. *Archives of Neurology*, 60(2):203–208, February 2003.

[75] Thomas Lumley. *Complex Surveys: A Guide to Analysis Using R.* Wiley, Hoboken, New Jersey, March 2010.

[76] Thomas Lumley and Alastair Scott. AIC and BIC for modeling with complex survey data. *Journal of Survey Statistics and Methodology*, 3(1):1–18, March 2015.

[77] Thomas Lumley and Alastair Scott. Fitting Regression Models to Survey Data. *Statistical Science*, 32(2):265–278, May 2017.

[78] C. L. Mallows. Some Comments on CP. *Technometrics*, 15(4):661–675, 1973. Publisher: [Taylor & Francis, Ltd., American Statistical Association, American Society for Quality].

[79] K. H. Masaki, K. G. Losonczy, G. Izmirlian, D. J. Foley, G. W. Ross, H. Petrovitch, R. Havlik, and L. R. White. Association of vitamin E and C supplement use with cognitive function and dementia in elderly men. *Neurology*, 54(6):1265–1272, March 2000.

[80] Matthew A. Masten and Alexandre Poirier. Identification of Treatment Effects Under Conditional Partial Independence. *Econometrica*, 86(1):317–351, 2018. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.3982/ECTA14481.

[81] Ann M Mayo. Use of the Functional Activities Questionnaire in Older Adults with Dementia. *Alzheimer's Association*, (D13), 2016.

[82] Guido Mazzinari, Ary Serpa Neto, Sabrine N. T. Hemmes, Goran Hedenstierna, Samir Jaber, Michael Hiesmayr, Markus W. Hollmann, Gary H. Mills, Marcos F. Vidal Melo, Rupert M. Pearse, Christian Putensen, Werner Schmid, Paolo Severgnini, Hermann Wrigge, Oscar Diaz Cambronero, Lorenzo Ball, Marcelo Gama de Abreu, Paolo Pelosi, Marcus J. Schultz, Wolfgang Kroell, Helfried Metzler, Gerd Struber, Thomas Wegscheider, Hans Gombotz, Michael Hiesmayr, Werner Schmid, Bernhard Urbanek, David Kahn, Mona Momeni, Audrey Pospiech, Fernande Lois, Patrice Forget, Irina Grosu, Jan Poelaert, Veerle van Mossevelde, Marie-Claire van Malderen, Dimitri Dylst, Jeroen van Melkebeek, Maud Beran, Stefan de Hert, Luc De Baerdemaeker, Bjorn Heyse, Jurgen Van Limmen, Piet Wyffels, Tom Jacobs, Nathalie Roels, Ann De Bruyne, Stijn van de Velde, Brigitte Leva, Sandrine Damster, Benoit Plichon, Marina Juros-Zovko, Dejana Djonoviċ-Omanoviċ, Selma Pernar, Josip Zunic, Petar Miskovic, Antonio Zilic, Slavica Kvolik, Dubravka Ivic, Darija Azenic-Venzera, Sonja Skiljic, Hrvoje Vinkovic, Ivana Oputric, Kazimir Juricic, Vedran Frkovic, Jasminka Kopic, Ivan Mirkovic, Nenad Karanovic, Mladen Carev, Natasa Dropulic, Jadranka Pavicic Saric, Gorjana Erceg, Matea Bogdanovic Dvorscak, Branka Mazul-Sunko, Anna Marija Pavicic, Tanja Goranovic, Branka Maldini, Tomislav Radocaj, Zeljka Gavranovic, Inga Mladic-Batinica, Mirna Sehovic, Petr Stourac, Hana Harazim, Olga Smekalova, Martina Kosinova, Tomas Kolacek, Kamil Hudacek, Michal Drab, Jan Brujevic, Katerina Vitkova, Katerina Jirmanova, Ivana Volfova, Paula Dzurnakova, Katarina Liskova, Radovan Dudas, Radek Filipsky, Samir el Kafrawy, Hisham Hosny Abdelwahab, Tarek Metwally, Ahmed Abdel-Razek, Ahmed Mostafa El-Shaarawy, Wael Fathy Hasan, Ahmed Gouda Ahmed, Hany Yassin, Mohamed Magdy, Mahdy Abdelhady, Mohamed Mahran, Eiko Herodes, Peeter Kivik, Juri Oganjan, Annika Aun, Alar Sormus, Kaili Sarapuu, Merilin Mall, Juri Karjagin, Emmanuel Futier, Antoine Petit, Adeline Gerard, Emmanuel Marret, Marc Solier, Samir Jaber, Albert Prades, Jens Krassler, Simone Merzky, Marcel Gama de Abreu, Christopher Uhlig, Thomas Kiss, Anette Bundy, Thomas Bluth, Andreas Gueldner, Peter Spieth, Martin Scharffenberg, Denny Tran Thiem, Thea Koch, Tanja Treschan, Maximilian Schaefer, Bea Bastin, Johann Geib, Martin Weiss, Peter Kienbaum, Benedikt Pannen, Andre Gottschalk, Mirja Konrad, Diana Westerheide, Ben Schwerdtfeger, Hermann Wrigge, Philipp Simon, Andreas Reske, Christian Nestler, Dimitrios Valsamidis, Konstantinos Stroumpoulis, Georgios Antholopoulos, Antonis Andreou, Dimitris Karapanos, Kassiani Theodoraki, Georgios Gkiokas, Marios-Konstantinos Tasoulis, Tatiana Sidiropoulou, Foteini Zafeiropoulou, Panagiota Florou, Aggeliki Pandazi, Georgia Tsaousi, Christos Nouris, Chryssa Pourzitaki, Dmitri Bystritski, Reuven Pizov, Arieh Eden, Caterina Valeria Pesce, Annamaria Campanile, Antonella Marrella, Salvatore Grasso, Michele De Michele, Francesco Bona, Gianmarco Giacoletto, Elena Sardo, Luigi Giancarlo, Vicari Sottosanti, Maurizio Solca, Carlo Alberto Volta, Savino Spadaro, Marco Verri, Riccardo Ragazzi, Roberto Zoppellari, Gilda Cinnella, Pasquale Raimondo, Daniela La Bella, Lucia Mirabella, Davide D'antini, Paolo Pelosi, Alexandre Molin, Iole Brunetti, Angelo Gratarola, Giulia Pellerano, Rosanna Sileo, Stefano Pezzatto, Luca Montagnani, Laura Pasin, Giovanni Landoni, Alberto Zangrillo, Luigi Beretta, Ambra Licia Di Parma, Valentina Tarzia, Roberto Dossi, Marta Eugenia Sassone, Daniele Sances, Stefano Tredici, Gianluca Spano, Gianluca Castellani, Luigi

Delunas, Sopio Peradze, Marco Venturino, Ines Arpino, Sara Sher, Concezione Tommasino, Francesca Rapido, Paola Morelli, Maria Vargas, Giuseppe Servillo, Andrea Cortegiani, Santi Maurizio Raineri, Francesca Montalto, Vincenzo Russotto, Antonino Giarratano, Marco Baciarello, Michela Generali, Giorgia Cerati, Yigal Leykin, Filippo Bressan, Vittoria Bartolini, Lucia Zamidei, Luca Brazzi, Corrado Liperi, Gabriele Sales, Laura Pistidda, Paolo Severgnini, Elisa Brugnoni, Giuseppe Musella, Alessandro Bacuzzi, Dalip Muhardri, Agreta Gecaj-Gashi, Fatos Sada, Adem Bytyqi, Aurika Karbonskiene, Ruta Aukstakalniene, Zivile Teberaite, Erika Salciute, Renatas Tikuisis, Povilas Miliauskas, Sipylaite Jurate, Egle Kontrimaviciute, Gabija Tomkute, John Xuereb, Maureen Bezzina, Francis Joseph Borg, Sabrine Hemmes, Marcus Schultz, Markus Hollmann, Irene Wiersma, Jan Binnekade, Lieuwe Bos, Christa Boer, Anne Duvekot, Bas in 't Veld, Alice Werger, Paul Dennesen, Charlotte Severijns, Jasper De Jong, Jens Hering, Rienk van Beek, Stefan Ivars, Ib Jammer, Alena Breidablik, Katharina Skirstad Hodt, Frode Fjellanger, Manuel Vico Avalos, Jannicke Mellin-Olsen, Elisabeth Andersson, Amir Shafi-Kabiri, Ruby Molina, Stanley Wutai, Erick Morais, Glória Tareco, Daniel Ferreira, Joana Amaral, Maria de Lurdes Goncalves Castro, Susana Cadilha, Sofia Appleton, Suzana Parente, Mariana Correia, Diogo Martins, Angela Monteirosa, Ana Ricardo, Sara Rodrigues, Lucian Horhota, Ioana Marina Grintescu, Liliana Mirea, Ioana Cristina Grintescu, Dan Corneci, Silvius Negoita, and for the LAS VEGAS study–investigators. The Association of Intraoperative driving pressure with postoperative pulmonary complications in open versus closed abdominal surgery patients – a posthoc propensity score–weighted cohort analysis of the LAS VEGAS study. *BMC Anesthesiology*, 21(1):84, March 2021.

[83] M. C. Morris, L. A. Beckett, P. A. Scherr, L. E. Hebert, D. A. Bennett, T. S. Field, and D. A. Evans. Vitamin E and vitamin C supplement use and risk of incident Alzheimer disease. *Alzheimer Disease and Associated Disorders*, 12(3):121–126, September 1998.

[84] Krista L. Moulder, Lilah M. Besser, Duane Beekly, Kaj Blennow, Walter Kukull, and John C. Morris. Factors Influencing Successful Lumbar Puncture in Alzheimer Research. *Alzheimer Disease and Associated Disorders*, 31(4):287–294, December 2017.

[85] J. A. Nelder and R. W. M. Wedderburn. Generalized Linear Models. *Journal of the Royal Statistical Society. Series A (General)*, 135(3):370–384, 1972. Publisher: [Royal Statistical Society, Wiley].

[86] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453, October 2019. Publisher: American Association for the Advancement of Science.

[87] Sam S. Oh, Joshua Galanter, Neeta Thakur, Maria Pino-Yanes, Nicolas E. Barcelo, Marquitta J. White, Danielle M. de Bruin, Ruth M. Greenblatt, Kirsten Bibbins-Domingo, Alan H. B. Wu, Luisa N. Borrell, Chris Gunter, Neil R. Powe, and Esteban G. Burchard. Diversity in Clinical and Biomedical Research: A Promise Yet to Be Fulfilled. *PLOS Medicine*, 12(12):e1001918, December 2015.

[88] Colm O'Muircheartaigh and Larry V. Hedges. Generalizing from unrepresentative experiments: a stratified propensity score approach. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 63(2):195–210, February 2014. Publisher: John Wiley & Sons, Ltd.

[89] David K. Park, Andrew Gelman, and Joseph Bafumi. Bayesian Multilevel Estimation with Poststratification: State-Level Estimates from National Polls. *Political Analysis*, 12(4):375–385, 2004. Publisher: Cambridge University Press.

[90] Ronald C. Petersen, Ronald G. Thomas, Michael Grundman, David Bennett, Rachelle Doody, Steven Ferris, Douglas Galasko, Shelia Jin, Jeffrey Kaye, Allan Levey, Eric Pfeiffer, Mary Sano, Christopher H. van Dyck, and Leon J. Thal. Vitamin E and Donepezil for the Treatment of Mild Cognitive Impairment. *New England Journal of Medicine*, 352(23):2379–88, 2005.

[91] R Core Team. R: A Language and Environment for Statistical Computing, 2021.

[92] J. N. K Rao, C. F. J Wu, and K. Yue. Some Recent Work On Resampling Methods For Complex Surveys. *Survey Methodology*, 18(2):209–217, December 1992.

[93] Greg Ridgeway, Stephanie Ann Kovalchik, Beth Ann Griffin, and Mohammed U. Kabeto. Propensity Score Analysis with Survey Weighted Data. *Journal of causal inference*, 3(2):237–249, September 2015.

[94] Michael W. Robbins, Bonnie Ghosh-Dastidar, and Rajeev Ramchand. Blending Probability and Nonprobability Samples with Applications to a Survey of Military Caregivers | Journal of Survey Statistics and Methodology | Oxford Academic. *Journal of Survey Statistics and Methodology*, November 2020.

[95] Paul R. Rosenbaum and Donald B. Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, April 1983.

[96] Geoff Rowe and David Binder. Can Survey Bootstrap Replicates Be Used for Cross-Validation? *Section on Survey Research Methods at the Joint Statistical Meetings*, pages 1430–1437, 2008.

[97] D. B. Rubin. Estimating causal effects from large data sets using propensity scores. *Annals of Internal Medicine*, 127(8 Pt 2):757–763, October 1997.

[98] Donald B. Rubin. *Multiple Imputation for Nonresponse in Surveys.* Wiley, 1987.

[99] Jonathan D. Rubright, Mark S. Cary, Jason H. Karlawish, and Scott Y. H. Kim. Measuring how people view biomedical research: Reliability and validity analysis of the Research Attitudes Questionnaire. *Journal of Empirical Research on Human Research Ethics*, 6(1):63–68, March 2011.

[100] Christian R. Salazar, Dan Hoang, Daniel L. Gillen, and Joshua D. Grill. Racial and ethnic differences in older adults' willingness to be contacted about Alzheimer's disease research participation. *Alzheimer's & Dementia: Translational Research & Clinical Interventions*, 6(1):e12023, January 2020.

[101] Mary Sano, Christopher Ernesto, Ronald G. Thomas, Melville R. Klauber, Kimberly Schafer, Michael Grundman, Peter Woodbury, John Growdon, Carl W. Cotman, Eric Pfeiffer, Lon S. Schneider, and Leon J. Thal. A Controlled Trial of Selegiline, Alpha-Tocopherol, or Both as Treatment for Alzheimer's Disease. *New England Journal of Medicine*, 336(17):1216–1222, April 1997. Publisher: Massachusetts Medical Society _eprint: https://doi.org/10.1056/NEJM199704243361704.

[102] Terry L. Schell, Samuel Peterson, Brian G. Vegetabile, Adam Scherling, Rosanna Smart, and Andrew R. Morral. State-Level Estimates of Household Firearm Ownership. April 2020. Publisher: RAND Corporation.

[103] Jonathan S. Schildcrout and Paul J. Rathouz. Longitudinal studies of binary response data following case-control and stratified case-control sampling: design and analysis. *Biometrics*, 66(2):365–373, June 2010.

[104] Gideon Schwarz. Estimating the Dimension of a Model. *The Annals of Statistics*, 6(2):461–464, March 1978. Publisher: Institute of Mathematical Statistics.

[105] Stephen Senn, Erika Graf, and Angelika Caputo. Stratification for the propensity score compared with linear regression techniques to assess the effect of treatment or exposure. *Statistics in Medicine*, 26(30):5529–5544, 2007. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/sim.3133.

[106] M. Stone. Cross-Validatory Choice and Assessment of Statistical Predictions. *Journal of the Royal Statistical Society. Series B (Methodological)*, 36(2):111–147, 1974. Publisher: [Royal Statistical Society, Wiley].

[107] M. Stone. An Asymptotic Equivalence of Choice of Model by Cross-Validation and Akaike's Criterion. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):44–47, 1977. Publisher: [Royal Statistical Society, Wiley].

[108] Elizabeth A. Stuart. Matching methods for causal inference: A review and a look forward. *Statistical science : a review journal of the Institute of Mathematical Statistics*, 25(1):1–21, February 2010.

[109] University of Wisconsin School of Medicine and Public Health. Area Deprivation Index, 2021.

[110] Brian G. Vegetabile, Beth Ann Griffin, Donna L. Coffman, Matthew Cefalu, Michael W. Robbins, and Daniel F. McCaffrey. Nonparametric estimation of population average dose-response curves using entropy balancing weights for continuous exposures. *Health Services and Outcomes Research Methodology*, 21(1):69–110, March 2021.

[111] Grace Wahba. Spline bases, regularization, and generalized cross-validation for solving approximation problems with large quantities of noisy data. *Proceedings of the International Conference on Approximation theory in Honour of George Lorenz*, pages 905–912, 1980.

[112] Sally P. Walsh, Rema Raman, Katherine B. Jones, Paul S. Aisen, and Alzheimer's Disease Cooperative Study Group. ADCS Prevention Instrument Project: the Mail-In Cognitive Function Screening Instrument (MCFSI). *Alzheimer Disease and Associated Disorders*, 20(4 Suppl 3):S170–178, December 2006.

[113] Lingxiao Wang, Richard Valliant, and Yan Li. Adjusted Logistic Propensity Weighting Methods for Population Inference using Nonprobability Volunteer-Based Epidemiologic Cohorts. *arXiv:2007.02476 [stat]*, February 2021. arXiv: 2007.02476.

[114] Sherry Weitzen, Kate L. Lapane, Alicia Y. Toledano, Anne L. Hume, and Vincent Mor. Principles for modeling propensity scores in medical research: a systematic literature review. *Pharmacoepidemiology and Drug Safety*, 13(12):841–853, 2004. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/pds.969.

[115] Jerzy Wieczorek, Cole Guerin, and Thomas McMahon. K-fold cross-validation for complex sample surveys. *Stat*, 11(1):e454, 2022. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/sta4.454.

[116] Jiali Yu, Michael D. Green, Shasha Li, Yilun Sun, Sara N. Journey, Jae Eun Choi, Syed Monem Rizvi, Angel Qin, Jessica J. Waninger, Xueting Lang, Zoey Chopra, Issam El Naqa, Jiajia Zhou, Yingjie Bian, Long Jiang, Alangoya Tezel, Jeremy Skvarce, Rohan K. Achar, Merna Sitto, Benjamin S. Rosen, Fengyun Su, Sathiya P. Narayanan, Xuhong Cao, Shuang Wei, Wojciech Szeliga, Linda Vatan, Charles Mayo, Meredith A. Morgan, Caitlin A. Schonewolf, Kyle Cuneo, Ilona Kryczek, Vincent T. Ma, Christopher D. Lao, Theodore S. Lawrence, Nithya Ramnath, Fei Wen, Arul M. Chinnaiyan, Marcin Cieslik, Ajjai Alva, and Weiping Zou. Liver metastasis restrains immunotherapy efficacy via macrophage-mediated T cell elimination. *Nature Medicine*, 27(1):152–164, January 2021. Bandiera_abtest: a Cg_type: Nature Research Journals Number: 1 Primary_atype: Research Publisher: Nature Publishing Group Subject_term: Cancer immunotherapy;Metastasis;Monocytes and macrophages;T cells;Tumour immunology Subject_term_id: cancer-immunotherapy;metastases;monocytes-and-macrophages;t-cells;tumour-immunology.

[117] Bianca Zadrozny. Learning and evaluating classifiers under sample selection bias. *Proceedings, Twenty-First International Conference on Machine Learning, ICML 2004*, pages 903–910, 2004.

[118] Peter P. Zandi, James C. Anthony, Ara S. Khachaturian, Stephanie V. Stone, Deborah Gustafson, JoAnn T. Tschanz, Maria C. Norton, Kathleen A. Welsh-Bohmer, John C. S. Breitner, and for the Cache County Study Group. Reduced Risk of Alzheimer Disease in Users of Antioxidant Vitamin Supplements: The Cache County Study. *Archives of Neurology*, 61(1):82–88, January 2004.

[119] Elaine L. Zanutto. A Comparison of Propensity Score and Linear Regression Analysis of Complex Survey Data. *Journal of Data Science*, 4(1):67–91, 2006. Publisher: School of Statistics, Renmin University of China.

Appendix A

Derivation of Variance Estimator for Chapter 3

In this section we derive the various components of I and Q introduced in Chapter 3 Section 3.2.4. Please note that the notation in this section follows the notation used in Chapter 3. Using iterated expectations we can show the cross-term R is

$$R = E_{\mathcal{P}}\Big[\sum_{i\in\mathcal{C}} \overline{U}_i \Big(\sum_{i\in\mathcal{C}} T_i^T + \sum_{i\in\mathcal{R}} T_i^T)\Big)\Big|Z = z\Big] = E_{\mathcal{P}}\Big[\sum_{i\in\mathcal{C}} \overline{U}_i \sum_{i\in\mathcal{C}} T_i^T \Big| Z = z\Big].$$

So the method of moments estimator for Q is

$$\widehat{Q} = \begin{pmatrix} TT^T & \widehat{R}^T \\ \widehat{R} & \overline{U}\,\overline{U}^T \end{pmatrix}\Bigg|_{\binom{\beta}{\gamma}=\binom{\widehat{\beta}}{\widehat{\gamma}}}$$

with $\widehat{R} = \sum_{i\in\mathcal{C}} \overline{U}_i \sum_{i\in\mathcal{C}} T_i^T$.

Now consider the the terms of I. I_{TT} is the derivative of T with respect to γ which is Fisher's information matrix for logistic regression,

$$I_{TT} = \sum_{i\in\mathcal{C}\cup\mathcal{R}} E_{\mathcal{P}}\Big(-\frac{\partial T_i}{\partial\gamma}\Big|X_i = x_i\Big) = \sum_{i\in\mathcal{C}\cup\mathcal{R}}\Big(x_i^T(P_{Ci}(1-P_{Ci}))x_i\Big).$$

I_{UU} is similar,

$$I_{UU} = -\sum_{i\in\mathcal{C}} E_{\mathcal{P}}\Big[w_{Ci}\frac{\partial U_j}{\partial\beta_k}\Big|Z = z\Big] = z^T M(\beta) z$$

where

$$M(\beta) = diag\Big(\frac{w_{Ci}(\partial\eta_i/\partial\mu_i)^{-2}}{V(\mu_i|Z_i = z_i)}\Big).$$

Finally, I_{UT} is

$$I_{UTjk} = -\sum_{i\in\mathcal{C}\cup\mathcal{R}} E_{\mathcal{P}}\Big[\frac{\partial U_{ki}}{\partial\gamma_j}\Big|Z_i = z_i\Big] = \sum_{i\in\mathcal{C}\cup\mathcal{R}} E_{\mathcal{P}}\Big[w_{Ci}\Big(\frac{\partial\mu_i}{\partial\beta_j}\Big)\Big(\frac{Y_i-\mu_i}{V(\mu_i)}\Big)X_{ij}\Big|Z_i = z_i\Big].$$

This term does not equal zero, because X_{ij} is not fixed. We can rearrange $I^{-1}QI^{-1}$ to point out the relationship to the design based variance used in the survey sampling literature ([77]). Define $\widehat{A} = \widehat{I}_{UU}$ and $\widehat{B} = \overline{U}\overline{U}^T$ so that,

$$\widehat{Q} = \begin{pmatrix} T \\ \overline{U} \end{pmatrix} \begin{pmatrix} T^T\overline{U}^T \end{pmatrix} = \begin{pmatrix} TT^T & \widehat{R}^T \\ \widehat{R} & \widehat{B} \end{pmatrix}.$$

Using the formula for blockwise inversion (Fact 2.17.1 in [14]),

$$I^{-1} = \begin{pmatrix} I_{TT} & 0 \\ I_{UT} & A \end{pmatrix}^{-1} = \begin{pmatrix} I_{TT}^{-1} & 0 \\ -A^{-1}I_{UT}I_{TT}^{-1} & A^{-1} \end{pmatrix}.$$

Combining I^{-1} and Q we obtain the proposed variance estimator,

$$\widehat{V}_{Prop}(\widehat{\beta}) = \widehat{A}^{-1}\widehat{B}\widehat{A}^{-1} - \widehat{A}^{-1}\widehat{I}_{UT}\widehat{I}_{TT}^{-1}\widehat{R}^{T}\widehat{A}^{-1}.$$

Appendix B

Bias Adjusted C2C Results

Table B.1: Bias adjusted C2C results: Odds Ratios and 95% confidence intervals are presented for the models from [100] assessing the relationship between race/ethnicity and 9 responses with adjustment variables. The models were fit without any propensity weights and with propensity weights estimated using logistic regression, covariate balancing propensity score (CBPS), entropy balancing (EB), and random forest (RF) methods.

Trial type	Model	Hispanic	NH Asian	NH Black
Physical Activity / Diet Modification	Unweighted	1.06 (0.52, 2.16)	0.68 (0.34, 1.35)	1.87 (0.25, 13.95)
	Logistic	1.52 (0.40, 5.83)	0.82 (0.23, 2.97)	4.54 (0.45, 45.38)
	CBPS	1.52 (0.40, 5.76)	0.83 (0.23, 3.02)	4.64 (0.46, 46.28)
	EB	4.80 (0.54, 42.33)	0.35 (0.06, 2.18)	4.61 (0.31, 69.07)
	RF	0.74 (0.16, 3.50)	0.87 (0.19, 4.09)	3.93 (0.36, 43.02)
Cognitive Testing	Unweighted	0.50 (0.22, 1.12)	0.52 (0.18, 1.52)	0.71 (0.09, 5.55)
	Logistic	0.73 (0.13, 4.26)	0.47 (0.06, 3.89)	0.95 (0.09, 10.14)
	CBPS	0.72 (0.13, 4.04)	0.45 (0.05, 3.72)	0.99 (0.09, 10.66)
	EB	1.75 (0.09, 33.19)	0.91 (0.03, 23.98)	0.93 (0.08, 11.60)
	RF	0.27 (0.06, 1.27)	0.88 (0.15, 5.36)	1.41 (0.13, 15.37)
MRI Scans	Unweighted	0.67 (0.39, 1.15)	1.34 (0.58, 3.12)	0.28 (0.12, 0.66)
	Logistic	2.73 (0.68, 10.90)	1.71 (0.47, 6.22)	0.10 (0.03, 0.30)
	CBPS	2.61 (0.67, 10.25)	1.70 (0.47, 6.17)	0.10 (0.03, 0.31)
	EB	2.99 (0.28, 31.66)	3.06 (0.59, 15.86)	0.06 (0.02, 0.23)
	RF	1.09 (0.28, 4.19)	1.98 (0.42, 9.35)	0.44 (0.08, 2.47)
PET Scans	Unweighted	0.66 (0.45, 0.97)	0.78 (0.49, 1.25)	0.46 (0.22, 0.98)
	Logistic	0.73 (0.21, 2.57)	0.48 (0.21, 1.12)	0.13 (0.04, 0.47)
	CBPS	0.72 (0.21, 2.47)	0.49 (0.21, 1.11)	0.13 (0.04, 0.48)
	EB	1.55 (0.31, 7.65)	0.47 (0.13, 1.70)	0.06 (0.02, 0.21)
	RF	0.81 (0.34, 1.94)	0.52 (0.20, 1.31)	0.42 (0.10, 1.74)
Blood Draws	Unweighted	0.62 (0.35, 1.10)	0.31 (0.18, 0.53)	0.27 (0.11, 0.67)
	Logistic	1.62 (0.45, 5.75)	0.37 (0.15, 0.89)	0.70 (0.15, 3.20)
	CBPS	1.58 (0.45, 5.54)	0.37 (0.15, 0.88)	0.70 (0.16, 3.16)
	EB	4.17 (0.69, 25.15)	0.24 (0.05, 1.07)	0.61 (0.12, 3.01)
	RF	0.77 (0.17, 3.47)	0.16 (0.04, 0.60)	0.29 (0.05, 1.61)
Approved Medications	Unweighted	0.68 (0.42, 1.10)	0.61 (0.36, 1.01)	0.67 (0.25, 1.80)
	Logistic	0.17 (0.06, 0.49)	0.66 (0.29, 1.50)	0.72 (0.21, 2.40)
	CBPS	0.17 (0.06, 0.49)	0.65 (0.29, 1.49)	0.72 (0.21, 2.41)
	EB	0.13 (0.02, 0.79)	0.43 (0.08, 2.44)	0.81 (0.10, 6.69)
	RF	0.64 (0.19, 2.17)	0.44 (0.15, 1.29)	0.45 (0.14, 1.45)
Investigational Medications	Unweighted	0.62 (0.42, 0.90)	0.55 (0.36, 0.83)	0.52 (0.24, 1.11)
	Logistic	0.31 (0.10, 0.93)	0.33 (0.14, 0.77)	0.70 (0.25, 1.98)
	CBPS	0.31 (0.10, 0.93)	0.33 (0.14, 0.76)	0.71 (0.25, 1.98)
	EB	0.50 (0.07, 3.73)	0.29 (0.08, 1.04)	0.82 (0.25, 2.75)
	RF	0.67 (0.26, 1.67)	0.40 (0.15, 1.08)	0.36 (0.11, 1.15)
Lumbar Puncture	Unweighted	0.95 (0.70, 1.30)	1.82 (1.26, 2.63)	0.25 (0.10, 0.62)
	Logistic	0.82 (0.33, 2.02)	1.06 (0.55, 2.05)	0.14 (0.03, 0.62)
	CBPS	0.82 (0.34, 2.01)	1.05 (0.55, 2.03)	0.14 (0.03, 0.64)
	EB	1.45 (0.35, 5.97)	0.64 (0.21, 2.00)	0.09 (0.01, 0.62)
	RF	1.27 (0.64, 2.50)	1.09 (0.45, 2.64)	1.07 (0.30, 3.85)
Autopsy	Unweighted	0.83 (0.59, 1.16)	0.43 (0.30, 0.62)	0.30 (0.16, 0.59)
	Logistic	0.50 (0.19, 1.33)	0.28 (0.10, 0.76)	0.55 (0.16, 1.88)
	CBPS	0.51 (0.20, 1.35)	0.28 (0.10, 0.74)	0.57 (0.17, 1.90)
	EB	0.32 (0.09, 1.13)	0.27 (0.04, 1.82)	0.41 (0.10, 1.62)
	RF	1.07 (0.54, 2.10)	0.28 (0.12, 0.65)	1.12 (0.34, 3.70)

Appendix C

Derivation of Variance Estimator for Chapter 4

In this appendix we derive the form of Fisher's expected information matrix, I, introduced in Equation 4.10 for the variance estimator defined in Chapter 4 Section 4.2.3. To simplify notation throughout the derivation, we have omitted the hats on the estimated sampling weights and propensity scores.

To make the derivation easier to follow, we are repeating the model definitions from Section 4.2.2 for estimating the sampling weights, propensity scores, and the causal effect as well as the corresponding estimating equations used to estimate the model parameters. Recall that sampling weights w_i are a function of sampling probabilities p_i where

$$w_i \propto \frac{1 - p_i}{p_i}$$

and sampling probabilities are estimated using a logistic regression model with parameters

γ:

$$\Psi_i = \text{logit}(p_i) = v_i\gamma.$$

The model is fit by solving the following estimating equation:

$$T_m(\gamma) = \sum_{i \in \mathcal{C} \cup \mathcal{R}} T_{mi}(\gamma) = \sum_{i \in \mathcal{C} \cup \mathcal{R}} (C_i - p_i)v_{im} = 0.$$

Thus the estimated sampling weights w are a function of γ through the sampling probability p. The propensity scores $e(z)$ are estimated using a logistic regression model with parameters ξ,

$$\Phi_i = \text{logit}(e(z_i)) = z_i\xi,$$

and the model is fit by solving the weighted estimating equation,

$$\bar{S}_l(\xi, \gamma) = \sum_{i \in \mathcal{C}} \bar{S}_{li}(\xi, \gamma; \hat{w}) = \sum_{i \in \mathcal{C}} \hat{w}_i(a_i - e(z_i))z_{il} = 0.$$

Thus the estimated propensity scores $e(z)$ are functions of w, p, ξ, and γ. This fact will be useful when we take derivatives of the propensity score using the chain rule in the derivation. We then use the estimated sampling weights and propensity scores in the outcome model with parameter vector $\beta = [\beta_0 \quad \beta_1 \quad \beta_2]$ and covariates $x_i = [1 \quad a_i \quad \hat{e}(z_i)]$:

$$\eta_i = g(\mu_i) = \beta_0 + \beta_1 a_i + \beta_2 \hat{e}(z_i).$$

This model is fit with the weighted estimating equation:

$$\bar{U}_j(\beta, \xi, \gamma) = \sum_{i \in \mathcal{C}} \hat{w}_i U_{ji}(\beta, \xi, \gamma) = \sum_{i \in \mathcal{C}} \hat{w}_i \frac{(y_i - \mu_i)}{V(\mu_i)} \left[\frac{\partial \eta_i}{\partial \mu_i}\right]^{-1} x_{ij} = 0.$$

$$
\begin{aligned}
I_{TT} &= -\sum_{i \in \mathcal{C} \cup \mathcal{R}} E\Big[\frac{\partial T_i(\gamma)}{\partial \gamma}\Big] = \sum_{i \in \mathcal{C} \cup \mathcal{R}} v_i^T p_i(1-p_i) v_i \\
I_{SS} &= -\sum_{i \in \mathcal{C}} E\Big[\frac{\partial \bar{S}_i(\xi,\gamma)}{\partial \xi}\Big] = \sum_{i \in \mathcal{C}} w_i z_i^T e(x_i)(1-e(x_i)) z_i \\
I_{UU} &= -\sum_{i \in \mathcal{C}} E\Big[\frac{\partial \bar{U}_i(\beta,\xi,\gamma)}{\partial \beta}\Big] = \sum_{i \in \mathcal{C}} x_i^T \mathrm{diag}\Big[\frac{w_i}{V(\mu_i)}\Big\{\frac{\partial \eta_i}{\partial \mu_i}\Big\}^{-2}\Big] x_i
\end{aligned}
$$

I_{ST} requires the chain rule

$$
I_{STlk} = -\sum_{i \in \mathcal{C}} E\Big[\frac{\partial \bar{S}_i(\xi,\gamma)_l}{\partial \gamma_k}\Big] = -\sum_{i \in \mathcal{C}} E\Big[\frac{\partial \bar{S}_i(\xi,\gamma)_l}{\partial w_i}\frac{\partial w_i}{\partial p_i}\frac{\partial p_i}{\partial \Psi_i}\frac{\partial \Psi_i}{\partial \gamma_k}\Big]
$$

where

$$
\begin{aligned}
\frac{\partial \bar{S}_i(\xi,\gamma)_l}{\partial w_i} &= S_i(\xi,\gamma)_l \\
\frac{\partial w_i}{\partial p_i} &= -p_i^{-2} \\
\frac{p_i}{\partial \Psi_i} &= p_i(1-p_i) \\
\frac{\Psi_i}{\partial \gamma_k} &= v_{ik}
\end{aligned}
$$

Thus,

$$
\begin{aligned}
I_{STlk} &= -\sum_{i \in \mathcal{C}} E\Big[\frac{\partial \bar{S}_i(\xi,\gamma)_l}{\partial \gamma_k}\Big] \\
&= -\sum_{i \in \mathcal{C}} E\Big[(S_{il})(-p_i^{-2})p_i(1-p_i)v_{ik}\Big] \\
&= \sum_{i \in \mathcal{C}} E\Big[w_i(a_i - e(x_i))z_{il}v_{ik}\Big]
\end{aligned}
$$

Now, I_{US} requires the product rule and the chain rule

$$\begin{aligned}
I_{USjl} &= -\sum_{i\in\mathcal{C}} E\Big[\frac{\partial \bar{U}_i(\beta,\xi,\gamma)_j}{\partial \xi_l}\Big] \\
&= -\sum_{i\in\mathcal{C}} E\Bigg[\frac{\partial}{\partial \xi_l}\Big\{w_i \frac{y_i-\mu_i}{V(\mu_i)}\Big[\frac{\partial \eta_i}{\partial \mu_i}\Big]^{-1} x_{ij}\Big\}\Bigg]
\end{aligned}$$

Applying the product rule,

$$\begin{aligned}
I_{US} = -\sum_{i\in\mathcal{C}} E\Bigg[w_i\Bigg(&\frac{\partial}{\partial \xi_l}\Big\{y_i-\mu_i\Big\}\frac{1}{V(\mu_i)}\Big[\frac{\partial \eta_i}{\partial \mu_i}\Big]^{-1} x_{ij} + (y_i-\mu_i)\frac{\partial}{\partial \xi_l}\Big\{\frac{1}{V(\mu_i)}\Big\}\Big[\frac{\partial \eta_i}{\partial \mu_i}\Big]^{-1} x_{ij} \\
&+ \frac{y_i-\mu_i}{V(\mu_i)}\frac{\partial}{\partial \xi_l}\Big\{\Big[\frac{\partial \eta_i}{\partial \mu_i}\Big]^{-1}\Big\}x_{ij} + \frac{y_i-\mu_i}{V(\mu_i)}\Big[\frac{\partial \eta_i}{\partial \mu_i}\Big]^{-1}\frac{\partial}{\partial \xi_l}\Big\{x_{ij}\Big\}\Bigg)\Bigg]
\end{aligned}$$

Applying the chain rule,

$$\begin{aligned}
I_{USjl} = -\sum_{i\in\mathcal{C}} E\Bigg[w_i\Bigg(&\frac{\partial}{\partial \mu_i}\Big\{y_i-\mu_i\Big\}\frac{\partial \mu_i}{\partial \eta_i}\frac{\partial \eta_i}{\partial e(x_i)}\frac{\partial e(x_i)}{\partial \xi_l}\frac{1}{V(\mu_i)}\Big[\frac{\partial \eta_i}{\partial \mu_i}\Big]^{-1} x_{ij} \\
&+ (y_i-\mu_i)\frac{\partial}{\partial \mu_i}\Big\{\frac{1}{V(\mu_i)}\Big\}\frac{\partial \mu_i}{\partial \eta_i}\frac{\partial \eta_i}{\partial e(x_i)}\frac{\partial e(x_i)}{\partial \xi_l}\Big[\frac{\partial \eta_i}{\partial \mu_i}\Big]^{-1} x_{ij} \\
&+ \frac{y_i-\mu_i}{V(\mu_i)}\frac{\partial}{\partial \mu_i}\Big\{\Big[\frac{\partial \eta_i}{\partial \mu_i}\Big]^{-1}\Big\}\frac{\partial \mu_i}{\partial \eta_i}\frac{\partial \eta_i}{\partial e(x_i)}\frac{\partial e(x_i)}{\partial \xi_l}x_{ij} \\
&+ \frac{y_i-\mu_i}{V(\mu_i)}\Big[\frac{\partial \eta_i}{\partial \mu_i}\Big]^{-1}\frac{\partial x_{ij}}{\partial e(x_i)}\frac{\partial e(x_i)}{\partial \xi_l}\Bigg)\Bigg]\Bigg)\Bigg] \\
= -\sum_{i\in\mathcal{C}} E\Bigg[w_i\Bigg(&-\frac{\partial \eta_i}{\partial e(x_i)}\frac{\partial e(x_i)}{\partial \xi_l}\frac{1}{V(\mu_i)}\Big[\frac{\partial \eta_i}{\partial \mu_i}\Big]^{-2} x_{ij} \\
&+ \frac{-(y_i-\mu_i)}{V(\mu_i)^2}\frac{\partial}{\partial \mu_i}\Big\{V(\mu_i)\Big\}\frac{\partial \eta_i}{\partial e(x_i)}\frac{\partial e(x_i)}{\partial \xi_l}\Big[\frac{\partial \eta_i}{\partial \mu_i}\Big]^{-2} x_{ij} \\
&+ \frac{y_i-\mu_i}{V(\mu_i)}\frac{\partial}{\partial \mu_i}\Big\{\Big[\frac{\partial \eta_i}{\partial \mu_i}\Big]^{-1}\Big\}\Big[\frac{\partial \eta_i}{\partial \mu_i}\Big]^{-1}\frac{\partial \eta_i}{\partial e(x_i)}\frac{\partial e(x_i)}{\partial \xi_l}x_{ij} \\
&+ \frac{y_i-\mu_i}{V(\mu_i)}\Big[\frac{\partial \eta_i}{\partial \mu_i}\Big]^{-1}\frac{\partial x_{ij}}{\partial e(x_i)}\frac{\partial e(x_i)}{\partial \xi_l}\Bigg)\Bigg]
\end{aligned}$$

Pulling out common terms,

$$I_{USjl} = -\sum_{i\in\mathcal{C}} E\Bigg[w_i \frac{\partial e(x_i)}{\partial \xi_l}\Bigg(\frac{\partial \eta_i}{\partial e(x_i)} x_{ij}\Bigg[-\frac{1}{V(\mu_i)}\Big[\frac{\partial \eta_i}{\partial \mu_i}\Big]^{-2} + \frac{-(y_i-\mu_i)}{V(\mu_i)^2}\frac{\partial}{\partial \mu_i}\Big\{V(\mu_i)\Big\}\Big[\frac{\partial \eta_i}{\partial \mu_i}\Big]^{-2} + \frac{y_i-\mu_i}{V(\mu_i)}\frac{\partial}{\partial \mu_i}\Big\{\Big[\frac{\partial \eta_i}{\partial \mu_i}\Big]^{-1}\Big\}\Big[\frac{\partial \eta_i}{\partial \mu_i}\Big]^{-1}\Bigg] + \frac{y_i-\mu_i}{V(\mu_i)}\Big[\frac{\partial \eta_i}{\partial \mu_i}\Big]^{-1}\frac{\partial x_{ij}}{\partial e(x_i)}\Bigg)\Bigg]$$

Now consider the following partial derivatives,

$$\begin{aligned}
\frac{\partial e(x_i)}{\partial \xi_l} &= \frac{\partial e(x_i)}{\partial \Phi_i}\frac{\partial \Phi_i}{\partial \xi_l} = e(x_i)(1-e(x_i))z_{il} \\
\frac{\partial \eta_i}{\partial e(x_i)} &= \beta_2 \\
\frac{\partial x_{ij}}{\partial e(x_i)} &= I(j=3)
\end{aligned}$$

Plugging these terms in we get,

$$I_{USjl} = -\sum_{i\in\mathcal{C}} E\Bigg[w_i e(x_i)(1-e(x_i))z_{il}\Bigg(\beta_2 x_{ij}\Bigg[-\frac{1}{V(\mu_i)}\Big[\frac{\partial \eta_i}{\partial \mu_i}\Big]^{-2} + \frac{-(y_i-\mu_i)}{V(\mu_i)^2}\frac{\partial}{\partial \mu_i}\Big\{V(\mu_i)\Big\}\Big[\frac{\partial \eta_i}{\partial \mu_i}\Big]^{-2} + \frac{y_i-\mu_i}{V(\mu_i)}\frac{\partial}{\partial \mu_i}\Big\{\Big[\frac{\partial \eta_i}{\partial \mu_i}\Big]^{-1}\Big\}\Big[\frac{\partial \eta_i}{\partial \mu_i}\Big]^{-1}\Bigg] + I(j=3)\frac{y_i-\mu_i}{V(\mu_i)}\Big[\frac{\partial \eta_i}{\partial \mu_i}\Big]^{-1}\Bigg)\Bigg]$$

Putting it into matrix form we get,

$$I_{US} = -E\Bigg[X^T D_{US1} Z + \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} 1_{n_C}^T D_{US2} Z\Bigg]$$

where

$$
\begin{aligned}
D_{US1} &= \text{diag}\Bigg(w_i e(x_i)(1-e(x_i))\beta_2 \Bigg[-\frac{1}{V(\mu_i)}\Big[\frac{\partial \eta_i}{\partial \mu_i}\Big]^{-2} + \frac{-(y_i-\mu_i)}{V(\mu_i)^2}\frac{\partial}{\partial \mu_i}\Big\{V(\mu_i)\Big\}\Big[\frac{\partial \eta_i}{\partial \mu_i}\Big]^{-2} \\
&\quad + \frac{y_i-\mu_i}{V(\mu_i)}\frac{\partial}{\partial \mu_i}\Big\{\Big[\frac{\partial \eta_i}{\partial \mu_i}\Big]^{-1}\Big\}\Big[\frac{\partial \eta_i}{\partial \mu_i}\Big]^{-1}\Bigg]\Bigg) \\
D_{US2} &= \text{diag}\Bigg(w_i e(x_i)(1-e(x_i))\frac{y_i-\mu_i}{V(\mu_i)}\Big[\frac{\partial \eta_i}{\partial \mu_i}\Big]^{-1}\Bigg)
\end{aligned}
$$

To find I_{UT} we will also need to use the product rule and the quotient rule. Note that $\bar{U}(\beta, \xi, \gamma)$ is a function of γ through w_i, μ_i and x_{ij}.

$$
\begin{aligned}
I_{UTjk} &= -\sum_{i\in\mathcal{C}} E\Big[\frac{\partial \bar{U}_i(\beta,\xi,\gamma)_j}{\partial \gamma_k}\Big] \\
&= -\sum_{i\in\mathcal{C}} E\Bigg[\frac{\partial}{\partial \gamma_k}\Big\{w_i\frac{y_i-\mu_i}{V(\mu_i)}\Big[\frac{\partial \eta_i}{\partial \mu_i}\Big]^{-1}x_{ij}\Big\}\Bigg]
\end{aligned}
$$

Let's consider the following derivatives

$$
\begin{aligned}
\frac{\partial w_i}{\partial \gamma_k} &= \frac{\partial w_i}{\partial p_i}\frac{\partial p_i}{\partial \Psi_i}\frac{\partial \Psi_i}{\partial \gamma_k} = (-p_i^{-2})[p_i(1-p_i)](v_{ij}) = -w_i v_{ik} \\
\frac{\partial \mu_i}{\partial \gamma_k} &= \frac{\partial \mu_i}{\partial \eta_i}\frac{\partial \eta_i}{\partial e(x_i)}\frac{\partial e(x_i)}{\partial w_i}\frac{\partial w_i}{\partial \gamma_k} = \Big[\frac{\partial \eta_i}{\partial \mu_i}\Big]^{-1}\beta_2\frac{\partial e(x_i)}{\partial w_i}\frac{\partial w_i}{\partial \gamma_k} \\
\frac{\partial x_{ij}}{\partial \gamma_k} &= \frac{\partial x_{ij}}{\partial e(x_i)}\frac{\partial e(x_i)}{\partial w_i}\frac{\partial w_i}{\partial \gamma_k} = I(j=3)\frac{\partial e(x_i)}{\partial w_i}\frac{\partial w_i}{\partial \gamma_k}
\end{aligned}
$$

Now, $e(x_i)$ is a function of γ_k through w_i through the estimating equation $\bar{S}_l$,

$$
\bar{S}_l = \sum_{i\in\mathcal{C}} w_i(a_i - e(x_i))z_{il} = \sum_{i\in\mathcal{C}} w_i a_i z_{il} - \sum_{i\in\mathcal{C}} w_i e(x_i) z_{il}
$$

Applying the chain rule,

$$
\begin{aligned}
\frac{\partial e(x_i)}{\partial w_i} &= \frac{\partial e(x_i)}{\partial \bar{S}_l}\frac{\partial \bar{S}_l}{\partial w_i} \\
&= \left[\frac{\partial \bar{S}_l}{\partial e(x_i)}\right]^{-1}\frac{\partial \bar{S}_l}{\partial w_i} \\
&= \big[-w_i z_{il}\big]^{-1}\big((a_i - e(x_i))z_{il}\big) \\
&= -\frac{1}{w_i}(a_i - e(x_i))
\end{aligned}
$$

Applying the product rule,

$$
\begin{aligned}
I_{UTjk} = -\sum_{i\in\mathcal{C}} E\Bigg[&\frac{\partial w_i}{\partial \gamma_k}\frac{y_i-\mu_i}{V(\mu_i)}\left[\frac{\partial \eta_i}{\partial \mu_i}\right]^{-1}x_{ij} + w_i\frac{\partial}{\partial \gamma_k}\Big\{y_i - \mu_i\Big\}\frac{1}{V(\mu_i)}\left[\frac{\partial \eta_i}{\partial \mu_i}\right]^{-1}x_{ij} \\
&+ w_i(y_i-\mu_i)\frac{\partial}{\partial \gamma_k}\left\{\frac{1}{V(\mu_i)}\right\}\left[\frac{\partial \eta_i}{\partial \mu_i}\right]^{-1}x_{ij} \\
&+ w_i\frac{y_i-\mu_i}{V(\mu_i)}\frac{\partial}{\partial \gamma_k}\left\{\left[\frac{\partial \eta_i}{\partial \mu_i}\right]^{-1}\right\}x_{ij} + w_i\frac{y_i-\mu_i}{V(\mu_i)}\left[\frac{\partial \eta_i}{\partial \mu_i}\right]^{-1}\frac{\partial x_{ij}}{\partial \gamma_k}\Bigg]
\end{aligned}
$$

Applying the chain rule,

$$
\begin{aligned}
I_{UTjk} = -\sum_{i\in\mathcal{C}} E\Bigg[&\frac{\partial w_i}{\partial \gamma_k}\frac{y_i-\mu_i}{V(\mu_i)}\left[\frac{\partial \eta_i}{\partial \mu_i}\right]^{-1}x_{ij} \\
&+ w_i(-1)\frac{\partial \mu_i}{\partial \eta_i}\frac{\partial \eta_i}{\partial e(x_i)}\frac{\partial e(x_i)}{\partial w_i}\frac{\partial w_i}{\partial \gamma_k}\frac{1}{V(\mu_i)}\left[\frac{\partial \eta_i}{\partial \mu_i}\right]^{-1}x_{ij} \\
&+ w_i(y_i-\mu_i)\frac{-1}{V(\mu_i)^{-2}}\frac{\partial}{\mu_i}\Big\{V(\mu_i)\Big\}\frac{\partial \mu_i}{\partial \eta_i}\frac{\partial \eta_i}{\partial e(x_i)}\frac{\partial e(x_i)}{\partial w_i}\frac{\partial w_i}{\partial \gamma_k}\left[\frac{\partial \eta_i}{\partial \mu_i}\right]^{-1}x_{ij} \\
&+ w_i\frac{y_i-\mu_i}{V(\mu_i)}\frac{\partial}{\partial \mu}\left\{\left[\frac{\partial \eta_i}{\partial \mu_i}\right]^{-1}\right\}\frac{\partial \mu_i}{\partial \eta_i}\frac{\partial \eta_i}{\partial e(x_i)}\frac{\partial e(x_i)}{\partial w_i}\frac{\partial w_i}{\partial \gamma_k}x_{ij} \\
&+ w_i\frac{y_i-\mu_i}{V(\mu_i)}\left[\frac{\partial \eta_i}{\partial \mu_i}\right]^{-1}\frac{\partial x_{ij}}{\partial e(x_i)}\frac{\partial e(x_i)}{\partial w_i}\frac{\partial w_i}{\partial \gamma_k}\Bigg]
\end{aligned}
$$

Rearranging and collecting common terms,

$$\begin{aligned}
I_{UTjk} = -\sum_{i\in\mathcal{C}} E\Bigg[&\frac{\partial w_i}{\partial \gamma_k}\frac{1}{V(\mu_i)}\Big[\frac{\partial \eta_i}{\partial \mu_i}\Big]^{-1}\Bigg((y_i-\mu_i)x_{ij} \\
&+ w_i\frac{\partial \eta_i}{\partial e(x_i)}\frac{\partial e(x_i)}{\partial w_i}x_{ij}\Big\{-\Big[\frac{\partial \eta_i}{\partial \mu_i}\Big]^{-1}\Big(1+\frac{(y_i-\mu_i)}{V(\mu_i)}\frac{\partial}{\mu_i}\Big\{V(\mu_i)\Big\}\Big) \\
&+ (y_i-\mu_i)\frac{\partial}{\partial \mu}\Big\{\Big[\frac{\partial \eta_i}{\partial \mu_i}\Big]^{-1}\Big\}\Big\} + I(j=3)w_i(y_i-\mu_i)\frac{\partial e(x_i)}{\partial w_i}\Bigg)\Bigg]
\end{aligned}$$

Plugging in values for known partial derivatives,

$$\begin{aligned}
I_{UTjk} = -\sum_{i\in\mathcal{C}} E\Bigg[&\frac{-w_i v_{ik}}{V(\mu_i)}\Big[\frac{\partial \eta_i}{\partial \mu_i}\Big]^{-1}\Bigg((y_i-\mu_i)x_{ij} \\
&- \beta_2(a_i-e(x_i))x_{ij}\Big\{-\Big[\frac{\partial \eta_i}{\partial \mu_i}\Big]^{-1}\Big(1+\frac{y_i-\mu_i}{V(\mu_i)}\frac{\partial}{\mu_i}\Big\{V(\mu_i)\Big\}\Big) \\
&+ (y_i-\mu_i)\frac{\partial}{\partial \mu}\Big\{\Big[\frac{\partial \eta_i}{\partial \mu_i}\Big]^{-1}\Big\}\Big\} - I(j=3)(y_i-\mu_i)(a_i-e(x_i))\Bigg)\Bigg]
\end{aligned}$$

Putting this into matrix form,

$$I_{UT} = E\left[X^T D_{UT1}V + \begin{pmatrix}0\\0\\1\end{pmatrix} 1_{n_C}^T D_{UT2}V\right]$$

where

$$\begin{aligned}
D_{UT1} &= \frac{w_i}{V(\mu_i)}\Big[\frac{\partial \eta_i}{\partial \mu_i}\Big]^{-1}\Bigg((y_i-\mu_i)-\beta_2(a_i-e(x_i))\Big[-\Big[\frac{\partial \eta_i}{\partial \mu_i}\Big]^{-1}\Big(1+\frac{y_i-\mu_i}{V(\mu_i)}\frac{\partial}{\mu_i}\Big\{V(\mu_i)\Big\}\Big) \\
&\quad + (y_i-\mu_i)\frac{\partial}{\partial \mu}\Big\{\Big[\frac{\partial \eta_i}{\partial \mu_i}\Big]^{-1}\Big\}\Big]\Bigg) \\
D_{UT2} &= -\frac{w_i}{V(\mu_i)}\Big[\frac{\partial \eta_i}{\partial \mu_i}\Big]^{-1}(y_i-\mu_i)(a_i-e(x_i))
\end{aligned}$$

To estimate I, we can replace Fisher's expected information with Fisher's observed information matrix and remove the expectations.