Propensity Score Methods for Estimating Causal Effects from Complex Survey Data

Dissertation

Presented in Partial Fulfillment of the Requirements for the Degree Doctor of Philosophy in the Graduate School of The Ohio State University

By

Robert D. Ashmead, M.S.

Graduate Program in Biostatistics

The Ohio State University

2014

Dissertation Committee:

Dr. Bo Lu, Advisor
Dr. Rebecca Andridge
Dr. Eloise Kaiziar
Abstract

Propensity score based adjustments are popular in analyzing observational data. To obtain valid causal estimates, we often assume that the sample is a simple random sample from the population of interest or that the treatment effect is homogeneous across the population. When data from surveys with complex design are used, ad-hoc adjustments to incorporate survey weights are often applied without rigorous justification. In this dissertation, we propose a super population framework, which includes a pair of potential outcomes for every unit in the population, to streamline the propensity score analysis for complex survey data. Based on the proposed framework, we develop propensity score stratification, weighting, and matching estimators along with a new class of hybrid estimators and corresponding variance estimators that adjust for survey design features. Additionally, we argue that in this context we should estimate the propensity scores by a weighted logistic regression using the sampling weights. Various estimators are compared in simulation studies that calculate the bias, mean-squared error, and coverage of the estimators. As the treatment effect becomes more heterogeneous, the gains of adjusting for the survey design increase. Lastly, we demonstrate the proposed methods using a real data example that estimates the effect of health insurance on self-rated health for adults in Ohio who may be eligible for tax credits to purchase medical insurance from the healthcare insurance exchange.
For my family and friends. Thank you for your love and support.
Acknowledgments

I would first like to thank my advisor, Dr. Lu, for his always helpful and steady guidance during the past few years. It has been a truly wonderful experience to learn from and collaborate with you. I would also like to acknowledge Dr. Andridge, Dr. Kaizar, and Dr. Stasny for their input, time, and encouragement as committee members. A special acknowledgment also goes to Dr. Nagaraja, who convinced me not give up on my PhD during my moment of crisis.

Thanks to everyone at the Government Resource Center, especially Tim Sahr, Rachel Tumin, and Dan Weston for their support, advice, and friendship during the past several years.

Thanks to my dear friends Dr. Chris and Robin Nau for your constant friendship, support, and countless Wednesday dinners over the past six years.

Finally, a profound thanks to my parents. Without your love, support, and encouragement, I would not be where I am today.
Vita

November 2, 1987 ....................... Born - Grand Rapids, Michigan

Education

2009 ........................................ B.A. Mathematics,
                                             College of Wooster,
                                             Wooster, Ohio

2011 ........................................ M.S. Statistics,
                                             The Ohio State University,
                                             Columbus, Ohio

Professional Experience

2009-2011 ................................. Graduate Teaching Associate,
                                             The Ohio State University,
                                             Columbus, Ohio

2011 ........................................ Summer Researcher,
                                             Eunice Kennedy Shriver National Institute
                                             of Child Health and Development,
                                             Bethesda, Maryland

2011-2012 ................................. Fellow,
                                             Leadership Education in Neurodevelopmental
                                             Disabilities,
                                             Columbus, Ohio

2012-Present ............................. Graduate Research Specialist,
                                             Ohio Colleges of Medicine Government
                                             Resource Center,
                                             Columbus, Ohio

Fields of Study

Major Field: Biostatistics
# Table of Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abstract</td>
<td>ii</td>
</tr>
<tr>
<td>Dedication</td>
<td>iii</td>
</tr>
<tr>
<td>Acknowledgments</td>
<td>iv</td>
</tr>
<tr>
<td>Vita</td>
<td>v</td>
</tr>
<tr>
<td>List of Tables</td>
<td>ix</td>
</tr>
<tr>
<td>List of Figures</td>
<td>x</td>
</tr>
<tr>
<td>1.  Introduction and Literature Review</td>
<td>1</td>
</tr>
<tr>
<td>1.1 The Potential Outcomes Framework</td>
<td>2</td>
</tr>
<tr>
<td>1.2 Population and Sample Average Treatment Effects</td>
<td>6</td>
</tr>
<tr>
<td>1.3 The Propensity Score</td>
<td>9</td>
</tr>
<tr>
<td>1.4 Common Propensity Score Methods</td>
<td>11</td>
</tr>
<tr>
<td>1.4.1 Propensity Score Matching</td>
<td>11</td>
</tr>
<tr>
<td>1.4.2 Propensity Score Stratification</td>
<td>14</td>
</tr>
<tr>
<td>1.4.3 Propensity Score Inverse Probability Weighting</td>
<td>16</td>
</tr>
<tr>
<td>1.5 Variance Estimators for Matching, Stratification, and IPW Estimators</td>
<td>21</td>
</tr>
<tr>
<td>1.5.1 Matching Estimators</td>
<td>21</td>
</tr>
<tr>
<td>1.5.2 Stratification Estimators</td>
<td>23</td>
</tr>
<tr>
<td>1.5.3 IPW Estimators</td>
<td>25</td>
</tr>
<tr>
<td>1.6 Complex Survey Sampling</td>
<td>31</td>
</tr>
<tr>
<td>1.7 Literature Review of Propensity Score Methods with Complex Survey Data</td>
<td>32</td>
</tr>
</tbody>
</table>
2. General Framework, Weighting and Stratification Estimators 37

2.1 A General Framework 37

2.1.1 Sampling Assumptions 40

2.2 Stratification Estimators 40

2.3 Weighting Estimators 45

2.4 Estimating the Propensity Score 54

2.5 Evaluating Balance 57

2.6 Variance Estimators 59

2.6.1 Variance of Stratification Estimators 59

2.6.2 Variance of Weighting Estimators 62

2.7 Estimators for $\Delta_{PATT}$ 66

2.7.1 Stratification Estimators 66

2.7.2 Weighting Estimators 67

2.8 Simulation Study 68

2.9 Tables and Figures 73

3. Stratification-Weighting Hybrid Estimators 78

3.1 Hybrid Estimators in the Non-Survey Context 79

3.1.1 Simulation 83

3.2 Hybrid Estimators in the Survey Context 88

3.2.1 Evaluating Balance 91

3.2.2 Simulation 92

3.3 Tables and Figures 94

4. Matching Estimators 99

4.1 Point Estimators 99

4.1.1 1:K Survey Matching Estimator 100

4.1.2 Survey Full Matching Estimator 102

4.2 Decomposition of the Estimators 104

4.3 Bias of the Estimators 105

4.4 Variance Estimators 106

4.5 Evaluating Balance 109

4.6 Simulation Studies 110

4.7 Tables and Figures 115

5. Example: Health Reform Impact on Individual Health Outcomes 119

5.1 Introduction 119

5.2 Data 120

5.3 Methods 120
5.4 Results ................................................................. 123
5.5 Conclusions .......................................................... 127
5.6 Tables and Figures .................................................. 129

6. Discussion and Future Work ........................................... 135
  6.1 Discussion ............................................................ 136
  6.2 Limitations ........................................................... 139
  6.3 Future Work ........................................................ 141
    6.3.1 Cluster Designs ............................................... 141
    6.3.2 Missing Data .................................................. 142
    6.3.3 Identifying Heterogeneity ................................... 143

Bibliography ............................................................... 144
## List of Tables

<table>
<thead>
<tr>
<th>Table</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1 Weighting and stratification simulation results with $\beta = \beta_{str.}$</td>
<td>76</td>
</tr>
<tr>
<td>2.2 Weighting and stratification simulation results with $\beta = \beta_{mod.}$</td>
<td>77</td>
</tr>
<tr>
<td>3.1 Mean number of propensity score strata used in propensity score stratification and hybrid estimators with stratification rules in a non-survey context.</td>
<td>97</td>
</tr>
<tr>
<td>3.2 Simulation results for propensity score stratification, weighting, and hybrid estimators when the propensity score model is misspecified in a non-survey context.</td>
<td>97</td>
</tr>
<tr>
<td>5.1 Point estimates, estimated variance, and estimated 95% confidence intervals for the effect of health insurance on self-rated health as a continuous variable.</td>
<td>133</td>
</tr>
<tr>
<td>5.2 Point estimates, estimated variance, and estimated 95% confidence intervals for the effect of health insurance on the prevalence of fair/poor self-rated health.</td>
<td>134</td>
</tr>
</tbody>
</table>
## List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>Graphical representation of the general potential outcomes framework.</td>
<td>73</td>
</tr>
<tr>
<td>2.2</td>
<td>Graphical representation of a typical potential outcomes framework.</td>
<td>73</td>
</tr>
<tr>
<td>2.3</td>
<td>Estimated propensity scores using weighted and unweighted logistic regression from one iteration of the simulation study.</td>
<td>74</td>
</tr>
<tr>
<td>2.4</td>
<td>Percent bias of propensity score stratification, weighting, and regression estimators.</td>
<td>74</td>
</tr>
<tr>
<td>2.5</td>
<td>Mean squared error of propensity score stratification, weighting, and regression estimators.</td>
<td>75</td>
</tr>
<tr>
<td>2.6</td>
<td>Coverage of the 95% confidence intervals of the propensity score stratification and weighting estimators.</td>
<td>75</td>
</tr>
<tr>
<td>3.1</td>
<td>Percent bias of propensity score stratification, weighting, and hybrid estimators with variable number of propensity score strata in a non-survey context.</td>
<td>94</td>
</tr>
<tr>
<td>3.2</td>
<td>MSE of propensity score stratification, weighting, and hybrid estimators with variable number of propensity score strata in a non-survey context.</td>
<td>95</td>
</tr>
<tr>
<td>3.3</td>
<td>MSE of propensity score stratification, weighting, hybrid, and regression estimators with stratification rules in a non-survey context.</td>
<td>95</td>
</tr>
<tr>
<td>3.4</td>
<td>Coverage of 95% confidence intervals created by propensity score stratification, weighting, and hybrid estimators with variable number of propensity score strata in a non-survey context.</td>
<td>96</td>
</tr>
</tbody>
</table>
3.5 Coverage of 95% confidence intervals created by propensity score stratification and hybrid estimators with stratification rules in a non-survey context. .................................................. 96

3.6 MSE of propensity score stratification, weighting, and hybrid estimators with variable number of propensity score strata in survey context with $\gamma = \gamma_{str}$. .................................................. 98

3.7 Coverage of 95% confidence intervals created by propensity score stratification, weighting, and hybrid estimators with variable number of propensity score strata in a survey context with $\gamma = \gamma_{str}$. .................................................. 98

4.1 Percent bias of propensity score weighting and matching estimators for $\Delta_{PATE}$. .................................................. 115

4.2 MSE of propensity score weighting and matching estimators for $\Delta_{PATE}$.116

4.3 Coverage of 95% confidence intervals created by propensity score weighting and matching estimators for $\Delta_{PATE}$. .................................................. 116

4.4 Histogram of the estimated propensity scores in a sample from one iteration of the $\Delta_{PATT}$ simulation. .................................................. 117

4.5 Percent bias and MSE of propensity score matching estimators for $\Delta_{PATT}$. .................................................. 117

4.6 Coverage of 95% confidence intervals created by propensity score matching estimators for $\Delta_{PATT}$. .................................................. 118

5.1 The smoothed distributions and box and whiskers plot of the estimated propensity score for treated and control groups. .................................................. 129

5.2 Estimated propensity scores using unweighted and weighted logistic regression. .................................................. 130

5.3 The balance of treated and control groups before and after propensity score weighting using weights implied by $\hat{\Delta}_{ipw2.svy}$, using propensity scores from weighted logistic regression. .................................................. 130

5.4 The balance of treated and control groups before and after propensity score weighting using weights implied by $\hat{\Delta}_{ipw2.svy}$, using propensity scores from an unweighted logistic regression. .................................................. 131
5.5 The balance of treated and control groups before and after stratification using $\hat{\Delta}_{s,svy}$. ................................................................. 131

5.6 The balance of treated and control groups before and after stratification and weighting using $\hat{\Delta}_{h,ipw2,svy}$. ................................................................. 132

5.7 The balance of treated and control groups before and after matching using $\hat{\Delta}_{matchK,svy}$. ................................................................. 132

5.8 The balance of treated and control groups before and after matching using $\hat{\Delta}_{matchfull,svy}$. ................................................................. 133
A major advantage of randomized studies over observational studies is that the randomization mechanism balances the distribution of the covariates across treatment groups. This includes both observed and unobserved covariates. In contrast, associations between the treatment exposure and the covariates may be present in an observational study leading to imbalance in the distribution of the covariates across treatment groups. If the covariates are also associated with the outcome, then without proper adjustment, confounding bias may occur when attempting to make inference about the treatment effect.

There exist many methods that attempt to adjust for confounding in observational data to get an unbiased estimate of the treatment effect, including matching on the covariates, stratifying on the covariates, weighting, and regression modeling. The propensity score, defined as the conditional probability of receiving treatment given observed covariates, was introduced by Rosenbaum and Rubin [36] and provides an approach to correcting the imbalance of observed covariates between treatment groups. The propensity score serves as a dimension reduction tool which summarizes treatment-assignment-related covariate information into a scalar value. Matching, stratifying, and weighting using the propensity score are popular methods for estimating treatment effects.
This dissertation develops new methods for estimating population average treatment effects using propensity scores when data is collected from a complex probability sample of a population. Specifically, methods based on weighting, stratifying, and matching by the estimated propensity score are examined. The remaining sections of this chapter give a review of the potential outcomes framework, propensity scores, propensity score estimators, complex sample surveys, and the existing literature relevant to our proposed methods. Next, Chapter 2 discusses our proposed general framework for potential outcomes in a complex probability survey, weighting and stratification estimators, and their respective variance estimators. We demonstrate the properties of these estimators in a simulation study. Chapter 3 proposes and investigates hybrid stratification-weighting estimators for both the complex survey and non-survey context. We illustrate how these estimators are more robust against misspecification of the propensity score through a simulation. In Chapter 4 we propose propensity score matching estimators for use with complex survey data and investigate their properties through a simulation. Chapter 5 consists of a real data example which uses our proposed methods to estimate the effect of health insurance on self-rated health. Lastly, in Chapter 6, we discuss the conclusions, limitations, and future directions of this research.

1.1 The Potential Outcomes Framework

The potential outcomes framework provides a way of thinking about the effect of a treatment or intervention on an individual, unit, sample, or population that is both intuitive and mathematically convenient. A treatment can be broadly defined depending on the area of study. Some examples are taking a drug (e.g. aspirin),
a health behavior (e.g. smoking cigarettes), or a program (e.g. having health insurance). The potential outcomes framework compares what the outcome would be under the treatment with what the outcome would be under some other treatment or lack of treatment. This comparison defines the treatment effect.

Consider a treatment or intervention with two levels of exposure, defined generically as treatment and control. Let $Z$ be an indicator of the observed treatment exposure, so that $Z = 1$ indicates receiving treatment and $Z = 0$ indicates receiving control. Define a pair of potential outcomes $(Y^z=1, Y^z=0)$ for each individual which represent the outcomes for that individual if he or she were to receive treatment ($Y^z=1$), and if he or she were to receive control ($Y^z=0$). The potential outcome framework is also known as counterfactual framework in psychological literature.

This framework allows us to define a causal effect for an individual. The treatment exposure has a causal effect for individual $i$ if $Y^z=1_i \neq Y^z=0_i$ [12]. The individual causal effect can be measured by comparisons of the potential outcomes such as $Y^z=1_i - Y^z=0_i$ or $Y^z=1_i / Y^z=0_i$. In real life we only can ever observe one of the potential outcomes for each individual, which leads to what Holland calls “the fundamental problem of causal inference” [17]. We are never able to “observe” the treatment effect for a unit because we only ever observe one potential outcome for each unit at the same time.

Instead of individual causal effects, we focus on average population causal effects. The treatment exposure has a causal effect in the population of $N$ units if $E[Y^z=1] \neq E[Y^z=0]$ [12], where

$$E[Y^z=1] = \frac{1}{N} \sum_{i=1}^{N} Y^z=1_i \quad \text{and} \quad E[Y^z=0] = \frac{1}{N} \sum_{i=1}^{N} Y^z=0_i$$

are the average of the potential outcomes of units in the population of interest. Here we borrow the notation of expectation to denote the population mean. From this point on we simply refer to the “average population causal effect” as the “average
causal effect” or “average treatment effect”. To quantify the strength of the causal effect we use effect measures such as

1. \( E[Y^{z=1}] - E[Y^{z=0}] \),

2. \( E[Y^{z=1}]/E[Y^{z=0}] \), or

3. \( (P[Y^{z=1} = 1])/P[Y^{z=1} = 0])/ (P[Y^{z=0} = 1])/P[Y^{z=0} = 0]) \)

depending on the type of outcome variable in our study and how we wish to interpret the effect [12]. Here we focus exclusively on estimating the average treatment effect difference \( E[Y^{z=1}] - E[Y^{z=0}] \). This measure is meaningful both when \( Y \) is a continuous variable, and when it is a binary variable.

Define the average treatment effect \( \Delta_{ATE} \) as

\[
\Delta_{ATE} = E[Y^{z=1}] - E[Y^{z=0}].
\] (1.1)

The average treatment effect \( \Delta_{ATE} \) can be thought of as the average difference in the outcome if all units received treatment rather than if all units had received control.

A slightly different estimand that is often used is the average treatment effect in the treated [6], which is defined as

\[
\Delta_{ATT} = E[Y^{z=1}|Z = 1] - E[Y^{z=0}|Z = 1].
\] (1.2)

The average treatment effect in the treated is essentially the treatment effect restricted to the treated subpopulation. Similarly, we can define the average treatment effect in the control

\[
\Delta_{ATC} = E[Y^{z=1}|Z = 0] - E[Y^{z=0}|Z = 0].
\] (1.3)
These estimands, $\Delta_{ATE}$, $\Delta_{ATT}$, and $\Delta_{ATC}$, will only differ when the treatment effect is heterogeneous. For example, assume that a binary variable $X$ is an effect modifier, and the distribution of $X$ differs between the treated and control populations. This implies that the distribution of $X$ is different between the overall, treated, and control populations, which implies that $\Delta_{ATE}$, $\Delta_{ATC}$, and $\Delta_{ATT}$ will differ. The choice of which estimand(s) is of interest depends on the particular research question and interpretation of the results of the study that the researcher is seeking. In general, in this dissertation we will focus on average treatment effects with a smaller amount of attention given to the average treatment effect in the treated. When using the treatment effect difference, such as in (1.1), the results for $\Delta_{ATT}$ apply directly to $\Delta_{ATC}$ because we can simply switch the meaning of treatment and control.

Assume we collect data, either experimental or observational, and define $Y$ as the observed outcome. We can write $Y$ as a combination of the potential outcomes and the observed treatment exposure, mainly

$$Y = Y^z=1Z + Y^z=0(1 - Z).$$

(1.4)

The problem in estimating (1.1) from the observed data is that while we may be able to observe $E[Y|Z = 1] = E[Y^z=1|Z = 1]$ and $E[Y|Z = 0] = E[Y^z=0|Z = 0]$, these are not necessarily equal to $E[Y^z=1]$ and $E[Y^z=0]$, respectively. In an observational study, $Z$ is not assigned randomly to individuals, and thus may be associated with the potential outcomes. In contrast, a randomized study guarantees that $(Y^z=1, Y^z=0) \perp \perp Z$, where $\perp \perp$ denotes statistical independence. Under this condition $E[Y|Z = 1] = E[Y^z=1]$ and $E[Y|Z = 0] = E[Y^z=0]$.

Now consider a set of covariates associated with both the potential outcomes and the treatment assignment. These covariates are typically called confounders. Let us
write these as \( \mathbf{X} \). If we are able to identify all confounders and measure them in \( \mathbf{X} \), then the potential outcomes are independent of the treatment assignment conditional on \( \mathbf{X} \). We write this condition as

\[
(Y^{z=1}, Y^{z=0}) \perp Z|\mathbf{X}. \tag{1.5}
\]

Condition (1.5) implies that among individuals with the same value of \( \mathbf{X} \), treatment assignment is independent of the potential outcomes. In other words, the treatment assignment is at random for those who share a value of \( \mathbf{X} \).

It then follows from (1.5) that \( E \{E[Y|Z = 1, \mathbf{X}]\} = E \{E[Y^{z=1}|\mathbf{X}]\} = E[Y^{z=1}] \), and \( E \{E[Y|Z = 0, \mathbf{X}]\} = E \{E[Y^{z=0}|\mathbf{X}]\} = E[Y^{z=0}] \), allowing us to estimate (1.1).

We emphasize that (1.5) is an un-testable assumption, and is commonly referred to as the assumption of strongly ignorable treatment assignment [36].

Similarly, we can get quantities needed for (1.2) and (1.3) by seeing that under condition (1.5)

\[
E_{\mathbf{X}|Z=0} \{E[Y|Z = 1, \mathbf{X}]\} = E_{\mathbf{X}|Z=0} \{E[Y^{z=1}|Z = 0, \mathbf{X}]\} = E[Y^{z=1}|Z = 0], \quad \text{and}
\]

\[
E_{\mathbf{X}|Z=1} \{E[Y|Z = 0, \mathbf{X}]\} = E_{\mathbf{X}|Z=1} \{E[Y^{z=0}|Z = 1, \mathbf{X}]\} = E[Y^{z=0}|Z = 1].
\]

### 1.2 Population and Sample Average Treatment Effects

Since we are ultimately interested in the complex probability survey setting, and not a simple random sample, it is important to consider differences between the sample and the population. One such difference to consider is the estimand. Consider a population of \( N \) units of which we sample \( n \). Let \( S \) be the set of \( n \) units in our sample and let \( i = 1, 2, \ldots, N \) index the population units. Define the population average treatment effect as
\[ \Delta_{PATE} = E[Y_{z=1}] - E[Y_{z=0}] = \frac{1}{N} \left( \sum_{i=1}^{N} Y_{i}^{z=1} - Y_{i}^{z=0} \right) \] (1.6)

which is the difference in potential outcomes averaged across the \( N \) units in the population [21]. This is the treatment effect in the population, rather than in our sample. The sample average treatment effect is defined as

\[ \Delta_{SATE} = E[Y_{z=1}|S = 1] - E[Y_{z=0}|S = 1] = \frac{1}{n} \left( \sum_{i \in S} Y_{i}^{z=1} - Y_{i}^{z=0} \right), \] (1.7)

which is the difference in potential outcomes averaged across the \( n \) units in the sample.

Similarly if we are concerned with average treatment effect in the treated we define

\[ \Delta_{PATT} = E[Y_{z=1}|Z = 1] - E[Y_{z=0}|Z = 1] \]
\[ = \frac{1}{N_1} \sum_{i=1}^{N_1} Z_i(Y_{i}^{z=1} - Y_{i}^{z=0}), \] and (1.8)

\[ \Delta_{SATT} = E[Y_{z=1}|Z = 1, S = 1] - E[Y_{z=0}|Z = 1, S = 1] \]
\[ = \frac{1}{n_1} \sum_{i \in S} Z_i(Y_{i}^{z=1} - Y_{i}^{z=0}), \] (1.9)

where \( N_1 \) and \( n_1 \) are the number of treated units in the population and sample respectively. For the average treatment effect in the control we define

\[ \Delta_{PATC} = E[Y_{z=1}|Z = 0] - E[Y_{z=0}|Z = 0] \]
\[ = \frac{1}{N_0} \sum_{i=1}^{N_0} (1 - Z_i)(Y_{i}^{z=1} - Y_{i}^{z=0}), \] and (1.10)

\[ \Delta_{SATC} = E[Y_{z=1}|Z = 0, S = 1] - E[Y_{z=0}|Z = 0, S = 1] \]
\[ = \frac{1}{n_0} \sum_{i \in S} (1 - Z_i)(Y_{i}^{z=1} - Y_{i}^{z=0}), \] (1.11)
where \( N_0 \) and \( n_0 \) are the number of control units in the population and sample respectively.

While the \( \Delta_{SATE} \), \( \Delta_{SATT} \), and \( \Delta_{SATC} \) are completely valid estimands and circumstances exist where they would be of interest, most of the time we are ultimately interested in estimating the corresponding population level quantities. When estimating \( \Delta_{PATE} \), \( \Delta_{PATT} \), or \( \Delta_{PATC} \) we worry about two situations that may result in bias: sample selection and treatment imbalance [21]. Treatment imbalance refers to differences in the distribution of confounders between treatment and control groups. Sample selection refers to differences in the distribution of the treatment effect between the sample and the population, meaning for example \( \Delta_{SATE} \neq \Delta_{PATE} \). Sample selection bias can only occur when we have a heterogeneous treatment effect. If the treatment effect is a function of a variable \( X \), and the distribution of \( X \) differs between our sample and population, then not correcting for the sample selection will lead to a biased estimate. Little and Rubin think about this idea in terms of a non-ignorable selection mechanism [25], meaning that the selection mechanism can be ignored if the sampled potential outcomes have the same distribution as the potential outcomes not in the sample. When attempting to estimate \( \Delta_{PATE} \), \( \Delta_{PATT} \), or \( \Delta_{PATC} \) with survey data, it is necessary to adjust for any differences between the sample and the population (sample selection) using the sampling weights and differences in covariate distributions of treated and control groups (treatment imbalance) using a method such as propensity scores.

In the context of a population-based sample survey, the assumption of a homogeneous treatment effect seems unwarranted in most cases, especially when dealing with human subjects. Kurth et al. [22] and Lunt et al. [28] show that the average
effect estimate is highly sensitive to the propensity score technique in the presence of
treatment effect heterogeneity because of the implied distribution of each method.

1.3 The Propensity Score

Propensity score methods are a useful way of summarizing confounders $X$ and can be used to estimate the causal effect of a treatment in observational studies. The propensity score is defined as the probability of treatment given the observed covariates. We write the propensity score as $e(X) = P(Z = 1|X)$. A necessary assumption is that the propensity score is bounded away from 0 and 1. Intuitively, if the propensity score is equal to zero or one for some observations, then it is difficult to argue that both potential outcomes exist.

The basis of all propensity score methodology is a result of the work of Rosenbaum and Rubin [36], who showed that under condition (1.5)

\[(Y^{z=1}, Y^{z=0}) \perp \perp Z | e(X) \text{ and } (1.12)\]

\[X \perp \perp Z | e(X). \text{ (1.13)}\]

Equation (1.12) implies that treatment assignment is at random among individuals who share the same propensity score. The major advantage of conditioning on the propensity score rather than $X$ is in dimensionality. The set of variables $X$ may potentially contain tens or hundreds of variables while $e(X)$ is only a single variable. It then follows from (1.12) that

\[E \{E[Y|Z = 1, e(X)]\} = E \{E[Y^{z=1}|e(X)]\} = E[Y^{z=1}], \text{ and}\]

\[E \{E[Y|Z = 0, e(X)]\} = E \{E[Y^{z=0}|e(X)]\} = E[Y^{z=0}],\]
allowing us to estimate the average treatment effect (1.1). Similarly following (1.12)

\[ E_{e(x)|Z=0} \{ E[Y|Z=1, e(x)] \} = E_{e(x)|Z=0} \{ E[Y^{z=1}|Z=0, e(x)] \} = E[Y^{z=1}|Z=0], \]

and

\[ E_{e(x)|Z=1} \{ E[Y|Z=0, e(x)] \} = E_{X|Z=1} \{ E[Y^{z=0}|Z=1, e(x)] \} = E[Y^{z=0}|Z=1], \]

which allows us to estimate (1.2) and (1.3).

A result of (1.13) is the distribution of the covariates \( X \) will be “balanced” in the sense of having equivalent distributions amongst those units in the treatment and control groups who share the same propensity score. In finite samples this balance is only approximate, and checking the balance is an important step in analysis to show that the propensity score is adequately modeled. Balance is a characteristic of our sample, not the population. However, analysis methods that we apply to our sample data may imply different distributions of \( X \) than the original sample through weighting observations, or simply choosing a matched subset of our sample. In these circumstances, the balance should be evaluated in these implied sample distributions rather than in the original sample.

The propensity score is generally not observed and therefore must be estimated from the observed data. Typically in practice we assume that \( e(X) \) follows a parametric model such as a logistic regression model or a probit regression model, though other modeling choices are possible. Let \( \hat{e} \) be the estimated propensity score from our model.
1.4 Common Propensity Score Methods

Three of the most common types of propensity score methods are matching, stratifying, and weighting with variations of each. Each method has its own advantages and disadvantages, though in practice matching seems to be the most commonly used.

1.4.1 Propensity Score Matching

Matching is a very appealing method because it is intuitive to non-researchers. We simply match units who are similar. Instead of matching treatment and control observations together based on one or several covariates, we create matched pairs or groups based on the estimated propensity score \( \hat{e} \). Once the matches are created, a number of analysis techniques can be used. While matched pairs or sets have the same or similar propensity scores, they do not necessarily have the exact same covariate pattern. Instead, the distribution of the covariates of entire set of matches should be balanced between the treatment and control groups.

Multiple matching designs are possible including 1:K matching, matching with replacement, and full matching. Assuming there are fewer treated units then control units, in 1:K matching, \( K \) control units are matched to each treated unit by the propensity score. Increasing \( K \) leads to a more efficient estimator given that the additional controls are good matches. Several methods can be used to estimate the treatment effect from the matched pairs or sets, the simplest being the mean of the pair/set differences which can be written as

\[
\hat{\Delta}_{\text{match}K} = \frac{1}{n_1} \sum_{j=1}^{n_1} Y_j^1 - \bar{Y}_j^0,
\]  

(1.14)
where \( j \) indexes the \( n_1 \) matched sets and \( \bar{Y}_j^0 \) is the average of the \( K \) control units in matched set \( j \) [7]. Not all control observations are included in (1.14), and as a result (1.14) estimates \( \Delta_{ATT} \) rather than \( \Delta_{ATE} \).

Matching with replacement attempts to utilize all observations in the data by matching each observation, treatment or control, with \( M \) observations of the opposite treatment assignment. The estimator can be written as

\[
\hat{\Delta}_{matchrep} = \frac{1}{n} \sum_{i=1}^{n} \left( \hat{Y}_i(1) - \hat{Y}_i(0) \right) = \frac{1}{n} \sum_{i=1}^{n} (2Z_i - 1) \left( 1 + \frac{K_M(i)}{M} \right) Y_i, \tag{1.15}
\]

where

\[
\hat{Y}_i(1) = \begin{cases} 
\frac{1}{M} \sum_{j \in J_M(i, \theta)} Y_j, & \text{if } Z_i = 0, \\
Y_i, & \text{if } Z_i = 1,
\end{cases}
\]

\[
\hat{Y}_i(0) = \begin{cases} 
Y_i, & \text{if } Z_i = 0, \\
\frac{1}{M} \sum_{j \in J_M(i, \theta)} Y_j, & \text{if } Z_i = 1,
\end{cases}
\]

\( J_M(i, \theta) \) is the set of \( M \) observations in the treatment group opposite to \( i \) with similar propensity scores, and \( K_M(i) \) is the number of times that unit \( i \) is used as a match [1]. This estimator, \( \hat{\Delta}_{matchrep} \), estimates \( \Delta_{ATE} \), but a simple modification can be made to estimate \( \Delta_{ATT} \) [3]. That estimator can be written as

\[
\hat{\Delta}^t_{matchrep} = \frac{1}{n_1} \sum_{Z_i=1} \left( Y_i - \hat{Y}_i(0) \right) = \frac{1}{n_1} \sum_{i=1}^{n_1} \left( Z_i - (1 - Z_i) \frac{K_M(i)}{M} \right) Y_i, \tag{1.16}
\]

Full matching uses all sampled units and creates matched sets which consist of one treated unit and multiple controls, or one control unit and multiple treated units with the number of units in a matched set being variable. This estimator can be written as
\[
\hat{\Delta}_{\text{match full}} = \frac{1}{n} \sum_{j=1}^{J} n_j \left( \frac{1}{\sum_{i=1}^{n_j} Z_{ji}} \sum_{i=1}^{n_j} Z_{ji} Y_{ji} - \frac{1}{\sum_{i=1}^{n_j} 1 - Z_{ji}} \sum_{i=1}^{n_j} (1 - Z_{ji}) Y_{ji} \right),
\]

(1.17)

where the data are divided into \( J \) mutually exclusive sets with similar propensity scores. \( Y_{ji} \) denotes the \( i \)th observation in the \( j \)th matched set. \( \hat{\Delta}_{\text{match full}} \) estimates \( \Delta_{ATE} \). Replacing the number of observations each matched set represents with the number of treated observations will instead give us an estimator of \( \Delta_{ATT} \). It can be written as

\[
\hat{\Delta}^t_{\text{match full}} = \frac{1}{n_1} \sum_{j=1}^{J} n_{1j} \left( \frac{1}{\sum_{i=1}^{n_j} Z_{ji}} \sum_{i=1}^{n_j} Z_{ji} Y_{ji} - \frac{1}{\sum_{i=1}^{n_j} 1 - Z_{ji}} \sum_{i=1}^{n_j} (1 - Z_{ji}) Y_{ji} \right),
\]

(1.18)

Rank-based tests like the Wilcoxon signed-rank test can be used to test for a treatment effect in matched samples as well. The Hodges-Lehmann estimator based on the signed-rank test can estimate the point estimate of the treatment effect.

In order to carry out any of these matching designs, we first must find good matches. A number of algorithms can be used to complete this task [38]. These include nearest neighbor matching, calipers, optimal matching, kernel matching, and combinations of each. In nearest neighbor matching units from the treated group are randomly ordered, and then in order, the control unit(s) with the nearest propensity score is chosen as a match. This is continued until all units are matched. Nearest neighbor matching can result in bad matches if there are no units with a similar propensity scores available. To guard against bad matches a caliper can be used which sets the maximum possible distance between two matches. If there is no control unit within the caliper of the treated unit, then the treated unit is not used in the analysis. Instead of finding the best match one at a time, optimal matching minimizes
an objective function across the entire matched set, such as the total distance between matches. It is more computationally expensive than nearest neighbor matching, but it should result in better matches. Kernel matching estimates the missing potential outcome for each unit in the treated group by taking a weighted average of nearby observations from the control group. The weights are based on the distances, the density function, and the bandwidth chosen.

### 1.4.2 Propensity Score Stratification

Stratification, also called subclassification, can be thought of as a coarser matching procedure. After estimation of the propensity score for all observations, the observations are split into $K$ strata based on their estimated propensity scores. The treatment effect is estimated within each stratum by treatment and control means and then averaged across strata. The estimator of $\Delta_{ATE}$ is given by

$$\hat{\Delta}_s = \sum_{k=1}^{K} \left( \frac{n_k}{n} \right) \left[ n_{1k}^{-1} \sum_{i=1}^{n} Z_i Y_i I(\hat{e}_i \in Q_k) - n_{0k}^{-1} \sum_{i=1}^{n} (1 - Z_i) Y_i I(\hat{e}_i \in Q_k) \right], \quad (1.19)$$

where $n$ is the sample size, $n_k$ is the sample size in stratum $k = 1, \ldots, K$, $n_{1k}$ and $n_{0k}$ are the number of treatment and control observations in stratum $k$, and $Q_k$ defines the $k$th stratum. If we are instead interested in estimating $\Delta_{ATT}$, replace $n_k$ with $n_{1k}$ and $n$ with $n_1$. This estimator can be written as

$$\hat{\Delta}_t = \sum_{k=1}^{K} \left( \frac{n_{1k}}{n_1} \right) \left[ n_{1k}^{-1} \sum_{i=1}^{n} Z_i Y_i I(\hat{e}_i \in Q_k) - n_{0k}^{-1} \sum_{i=1}^{n} (1 - Z_i) Y_i I(\hat{e}_i \in Q_k) \right], \quad (1.20)$$

Stratification works on the principle that within each stratum we have strongly ignorable treatment assignment, so that treatment assignment is at random within a stratum. This is typically only at best approximately true. In practice $K = 5$
strata defined by the quintiles of the estimated propensity score is often used due to a result by Rosenbaum and Rubin [37] that found stratifying on the quintiles eliminates approximately 90% of the bias. That result has been shown empirically to not be correct for all circumstances, even very reasonable ones [27]. In fact, the bias can be quite a lot higher, over 25% in some simulations [27].

Instead of stratifying on the quintiles, or other quantiles that create equal frequencies in strata, Myers and Louis investigate creating “optimal” stratifications, as a function of MSE [30]. They find that strata formed by “equal frequency” (quantiles of the propensity score) do as well or nearly as well as optimal stratification in almost all cases of their simulation. They conclude that the added analytical cost of calculating optimal strata, the additional estimation, and assumptions needed are not worth the small gains in MSE. They also recommend increasing the number of strata until the overall estimate remains relatively constant or until the estimated variance greatly increases.

The number of strata used has a bias-variance tradeoff relationship depending on the degree of confounding. In an un-confounded situation, the best solution in terms of MSE would be to have a single stratum, because this minimizes the variance. However, as the degree of confounding increases, having a larger number of strata to control the bias is the best solution. In their simulation Lunceford and Davidian found that increasing the number of strata from 5 to 10 reduced the bias by about 65% while keeping the variance roughly constant [27].

While creating strata strictly by “equal frequency” leaves us with the same number of total observations in each stratum, it does not guard against having too few treated or control observations in a stratum. This is a particular problem towards the extremes of the propensity score distribution. For instance, when the propensity
score is equal to 0.1 we expect 9 control observations for every 1 treated observation. This could lead to problems in the propensity score stratum that contains the observations with the smallest propensity scores if it is not large enough. There may be too few treated observations to properly estimate the variance, and additionally the few treated observations would be extremely influential in determining the estimated treated effect in that propensity score strata.

An extension of the stratification technique that seeks to reduce within-stratum confounding is to model the outcome through a regression model within each propensity score stratum and average the model-assisted estimated treatment effects across strata as before. The estimator can be written as

$$\hat{\Delta}_{sr} = \sum_{k=1}^{K} \left( \frac{n_k}{n} \right) \left[ n^{-1} \sum_{i=1}^{n} I(\hat{e}_i \in Q_k) \left\{ m^{(k)}(1, X_i, \hat{\alpha}^{(k)}) - m^{(k)}(0, X_i, \hat{\alpha}^{(k)}) \right\} \right], \quad (1.21)$$

where $m^{(k)}(Z, X, \hat{\alpha}^{(k)})$ is the regression outcome model in the $k$th stratum with estimated coefficients $\alpha^{(k)}$. $\hat{\Delta}_{sr}$ is a consistent estimator of $\Delta_{ATE}$ when we are able to correctly model the outcome, but adding the modeling component leaves us open to model misspecification [27]. The advantage of $\hat{\Delta}_{sr}$ over trying to directly model the outcome through a regression model is that $\hat{\Delta}_{sr}$ is less affected by model misspecification [27].

### 1.4.3 Propensity Score Inverse Probability Weighting

A third type of common propensity score method is inverse probability weighting (IPW). IPW uses the inverse of the propensity score as a weight similarly to the way the Horvitz-Thompson estimator uses the inverse probability of selection as a weight [19].
Recall equation (1.4), which defined \( Y = Y^{z=1}Z + Y^{z=0}(1 - Z) \). Also note that \( Z(1 - Z) = 0 \). These equations imply that \( E\left[ \frac{ZY}{e(X)} \right] = E\left[ \frac{Z^{Y^{z=1}}}{e(X)} \right] \), and assuming (1.12) we show that

\[
E\left[ \frac{ZY}{e(X)} \right] = E\left\{ E\left[ \frac{I(Z = 1)Y^{z=1}}{e(X)} \right] | Y^{z=1}, X \right\}
\]

\[
= E\left\{ \frac{Y^{z=1}}{e(X)} E\left[ I(Z = 1) \right] | Y^{z=1}, X \right\}
\]

\[
= E\left\{ \frac{Y^{z=1}}{e(X)} e(X) \right\}
\]

\[
= E\left[ Y^{z=1} \right].
\]

Similarly \( E\left[ \frac{(1-Z)Y}{(1-e(X))} \right] = E\left[ Y^{z=0} \right] \), which leads to an estimate of \( \Delta_{ATE} \) given by

\[
\hat{\Delta}_{ipw} = \frac{1}{n} \sum_{i=1}^{n} Z_i Y_i - \frac{1}{n} \sum_{i=1}^{n} (1 - Z_i) Y_i
\]

(1.22)

If we are instead interested in \( \Delta_{ATT} \), we can estimate it by

\[
\hat{\Delta}_{ipw} = \frac{1}{n_1} \sum_{i=1}^{n_1} Z_i Y_i - \frac{1}{n_1} \sum_{i=1}^{n_1} (1 - Z_i) Y_i \hat{e}_i
\]

(1.23)

Assuming (1.12) these weights give an unbiased estimator of \( \Delta_{ATT} \) because

\[
E[ZY] = E[ZY|Z = 1]P(Z = 1) + E[ZY|Z = 0]P(Z = 0) = E[Y^{z=1}|Z = 1]\left(\frac{n_1}{n}\right),
\]

and
\[
E \left[ \frac{(1 - Z)Ye(X)}{1 - e(X)} \right] = E \left\{ \frac{Y^{z=0}e(X)}{1 - e(X)} E \left[ 1 - I(Z = 1) | X, Y^{z=0} \right] \right\} \\
= E \left\{ Y^{z=0}e(X) \right\} \\
= E \left\{ E[Y^{z=0}|X, Z = 1] P(Z = 1|X) \right\} \\
= \int \left\{ E[Y^{z=0}|X, Z = 1] \right\} P(Z = 1|X) P(X = x) dx \\
= \int \left\{ \int Y^{z=0}P(Y^{z=0} = y|X, Z = 1) dy \right\} P(Z = 1|X) P(X = x) dx \\
= \int \left\{ \int Y^{z=0}P(Y^{z=0} = y|X, Z = 1) dy \right\} P(Z = 1, X = x) \frac{P(Z = 1)}{P(Z = 1)} dx \\
= P(Z = 1) \int \left\{ \int Y^{z=0}P(Y^{z=0} = y|X, Z = 1) dy \right\} P(X = x|Z = 1) dx \\
= P(Z = 1) \int \int Y^{z=0}P(Y^{z=0} = y, X, Z = 1) dy dx \\
= P(Z = 1) \int E[Y^{z=0}|X, Z = 1] P(X = x|Z = 1) dx \\
= \frac{n_1}{n} E[Y^{z=0}|Z = 1].
\]

A second version of the IPW estimator of \( \Delta_{ATE} \) is given by

\[
\hat{\Delta}_{ipw2} = \left( \sum_{i=1}^{n} \frac{Z_i}{\hat{e}_i} \right)^{-1} \sum_{i=1}^{n} \frac{Z_iY_i}{\hat{e}_i} - \left( \sum_{i=1}^{n} \frac{1 - Z_i}{1 - \hat{e}_i} \right)^{-1} \sum_{i=1}^{n} \frac{(1 - Z_i)Y_i}{(1 - \hat{e}_i)}, \quad (1.24)
\]

which divides the weighted sums of treated and control outcomes by the sum of their respective weights rather than by \( n \). This result is motivated by the fact that \( E[Z/e(X)] = E[(1 - Z)/(1 - e(X))] = 1 \), and performs better than (1.22) in practice [27].

The corresponding estimator of \( \Delta_{ATT} \) is given by by

\[
\hat{\Delta}^t_{ipw2} = (n_1)^{-1} \sum_{i=1}^{n} Z_iY_i - \left( \sum_{i=1}^{n} \frac{1 - Z_i}{1 - \hat{e}_i} \right)^{-1} \sum_{i=1}^{n} \frac{(1 - Z_i)Y_i}{(1 - \hat{e}_i)}, \quad (1.25)
\]

The estimators (1.22) and (1.24) fall in a more general class of estimators of \( \Delta_{ATE} \) defined by the solutions to the equations.
\[
\sum_{i=1}^{n} \left\{ \frac{Z_i(Y_i - \mu_1)}{e_i} + \eta_1 \left( \frac{Z_i - e_i}{e_i} \right) \right\} = 0, \text{ and }
\sum_{i=1}^{n} \left\{ \frac{(1 - Z_i)(Y_i - \mu_0)}{1 - e_i} - \eta_0 \left( \frac{Z_i - e_i}{1 - e_i} \right) \right\} = 0. \tag{1.26}
\]

Assuming the propensity score is known and setting \((\eta_0, \eta_1) = (\mu_0, \mu_1)\) gives \(\hat{\Delta}_{ipw1}\), while \((\eta_0, \eta_1) = (0, 0)\) yields \(\hat{\Delta}_{ipw2}\). While \(\hat{\Delta}_{ipw1}\) is an unbiased estimator of \(\Delta_{ATE}\), in general this class of estimators is not unbiased; however, they are consistent when the propensity score is correctly specified [33]. This can be seen from a combination of Slutsky’s theorem and the law of large numbers.

Finding the values \((\eta_1, \eta_0)\) that minimize the large sample variance for members of this class leads to \(\hat{\Delta}_{ipw3}\) [27], which is defined as

\[
\hat{\Delta}_{ipw3} = \left\{ \sum_{i=1}^{n} \frac{Z_i}{\hat{e}_i} \left( 1 - \frac{C_1}{\hat{e}_i} \right) \right\}^{-1} \sum_{i=1}^{n} \frac{Z_iY_i}{\hat{e}_i} \left( 1 - \frac{C_1}{\hat{e}_i} \right) - \left\{ \sum_{i=1}^{n} \frac{1 - Z_i}{1 - \hat{e}_i} \left( 1 - \frac{C_0}{1 - \hat{e}_i} \right) \right\}^{-1} \sum_{i=1}^{n} \frac{(1 - Z_i)Y_i}{(1 - \hat{e}_i)} \left( 1 - \frac{C_0}{1 - \hat{e}_i} \right), \tag{1.27}
\]

where

\[
C_1 = \frac{\sum_{i=1}^{n} \{(Z_i - \hat{e}_i)/\hat{e}_i\}}{\sum_{i=1}^{n} \{(Z_i - \hat{e}_i)/(1 - \hat{e}_i)\}^2}, \text{ and }
\]

\[
C_0 = -\frac{\sum_{i=1}^{n} \{(Z_i - \hat{e}_i)/(1 - \hat{e}_i)\}}{\sum_{i=1}^{n} \{(Z_i - \hat{e}_i)/(1 - \hat{e}_i)\}^2}.
\]

The terms \(1 - \frac{C_1}{\hat{e}_i}\) and \(1 - \frac{C_0}{1 - \hat{e}_i}\) reduce or increase the weight of each observation proportionately by how \(Z_i/\hat{e}_i\) deviates from its expectation of one. This results in numerical stability, especially with smaller sample sizes.

The more general class of estimators that estimate \(\Delta_{ATT}\) can be written as solutions to the equations

\[
19
\]
\[
\sum_{i=1}^{n} \{ Z_i(Y_i - \mu_1) \} = 0, \text{ and }
\sum_{i=1}^{n} \left\{ \frac{(1 - Z_i)(Y_i - \mu_0)e_i}{1 - e_i} - \eta_0 e_i \left( \frac{Z_i - e_i}{1 - e_i} \right) \right\} = 0, \tag{1.28}
\]
with \( \eta_0 = \mu_1 \) implying (1.23) and \( \eta_0 = 0 \) implying (1.25).

Minimizing the large sample variance with respect to \( \eta_0 \) yields the following estimator of \( \hat{\eta}_0 \)
\[
\hat{\eta}_0 = -E \left[ \frac{(1 - Z)(Y - \mu_0)e^2}{(1 - e)^2} \right] / E \left[ \frac{(Z - e)^2}{1 - e} e^2 \right].
\]

This leads to the estimator
\[
\hat{\Delta}_{ipw}^{t} = (n_1)^{-1} \sum_{i=1}^{n_1} Z_i Y_i - \left( \sum_{i=1}^{n_1} \frac{(1 - Z_i) \hat{e}_i}{1 - \hat{e}_i} \left( 1 - \frac{C_0 \hat{e}_i}{1 - \hat{e}_i} \right) \right)^{-1} \sum_{i=1}^{n_1} \frac{(1 - Z_i) \hat{e}_i Y_i}{(1 - \hat{e}_i)} \left( 1 - \frac{C_0 \hat{e}_i}{1 - \hat{e}_i} \right), \tag{1.29}
\]
where
\[
C_0 = -\frac{\sum_{i=1}^{n_1} \{(Z_i - \hat{e}_i) \hat{e}_i/(1 - \hat{e}_i)\}}{\sum_{i=1}^{n_1} \{(Z_i - \hat{e}_i) \hat{e}_i/(1 - \hat{e}_i)\}^2}.
\]

In general inverse probability weights can be thought of similarly to survey sampling weights in that each observed treated unit counts for itself and \( 1/(\hat{e}_i) - 1 \) other units not assigned to the treatment group, and each control unit counts for itself and \( 1/(1 - \hat{e}_i) - 1 \) other units not assigned to the control group when estimating \( \Delta_{ATE} \).
Likewise each observed treated unit counts for itself, and each control unit counts for itself and \( \hat{e}_i/(1 - \hat{e}_i) - 1 \) other units when estimating \( \Delta_{ATT} \). Inverse probability weighting seeks to create a pseudo-population in which the distribution of \( X \) is balanced between treatment and control units. When we are trying to estimate the average
treatment effect, this pseudo-population’s distribution of \( X \) should reflect the distribution of \( X \) in the whole sample. When estimating the average treatment effect in the treated, the pseudo-population’s distribution of \( X \) should reflect the distribution of \( X \) in the treated group.

Other weighting estimators are possible, many of which attempt to model the outcome parametrically. Robins et al. [33] propose a doubly robust estimator which utilizes a regression model of the outcome much in the same way as \( \hat{\Delta}_{sr} \). The doubly robust estimator of \( \Delta_{ATE} \) is defined as

\[
\hat{\Delta}_{dr} = \frac{1}{n} \sum_{i=1}^{n} \frac{Z_iY_i - (Z_i - \hat{e}_i)m_1(X_i, \hat{\alpha}_1)}{\hat{e}_i} - \frac{1}{n} \sum_{i=1}^{n} \frac{(1-Z_i)Y_i - (Z_i - \hat{e}_i)m_0(X_i, \hat{\alpha}_0)}{1 - \hat{e}_i},
\]

where \( m_1(X, \hat{\alpha}_1) \) and \( m_0(X, \hat{\alpha}_0) \) are the predicted outcomes for the treatment and control groups specified by regression models. When both the propensity score model and the outcome model are correctly specified, \( \hat{\Delta}_{dr} \) will have the smallest large sample variance of any weighting-type estimator [33]. It is also doubly robust in the sense that it is consistent if either the propensity score model or the outcome model is correctly specified.

1.5 Variance Estimators for Matching, Stratification, and IPW Estimators

1.5.1 Matching Estimators

Opinions on how to best estimate the standard error of matching estimators vary in the literature with the main argument being if we should treat our matches as individually matched sets or as overall matched groups [42] [15]. With the exclusion
of the non-parametric matching estimators, all the matching estimators discussed earlier can be motivated by either of these two philosophies.

When treated as fixed matches, the simplest variance estimation strategy is to estimate the variance of the matched set differences using the typical variance estimator, though matching with replacement creates problems with this approach. Additionally this approach does not take into account the variance due to the estimation of the propensity score or the matching process.

The motivation for treating the matches as matched groups instead of individuals is that matching on the propensity score does not guarantee matched units have the same covariate distribution. Instead, they should be approximately balanced overall. The matching process is a way to make comparable groups, and the actual matches are irrelevant. Instead of comparing group means of the matched sample groups, some authors recommend making further adjustments for confounders such as using a regression model to estimate the treatment effect [15]. This of course comes with the risk of model mis-specification, but also provides a way of estimating standard errors and a way to accommodate effect modifiers.

Lechner (2001) proposes the variance estimator for matching with replacement

\[ \hat{V}ar(\hat{\Delta}_{ATT}) \approx \frac{1}{N_1} \hat{V}ar(Y^z=1|Z = 1) + \frac{(\sum_{j \in I_0}(w_j)^2)}{(N_1)^2} \hat{V}ar(Y^z=0|Z = 0) \]  

(1.31)

where \( N_1 \) is the number of matched treated individuals and \( w_j \) is the number of times individual \( j \) from the control group has been matched [23].

Abadie and Imbens (2006) propose an estimator of the variance of their estimator \( \hat{\Delta}_{matchrep} \) (1.15), which is written as
\[ \text{Var}(\hat{\Delta}_{\text{matchrep}}) = \frac{1}{n^2} \sum_{i=1}^{n} (\hat{Y}_i^{z=1} - \hat{Y}_i^{z=0} - \hat{\Delta}_{\text{matchrep}})^2 \]
\[ + \frac{1}{n^2} \sum_{i=1}^{n} \left[ \left( \frac{K_M(i)}{M} \right)^2 + \left( \frac{2M - 1}{M} \right) \left( \frac{K_M(i)}{M} \right) \right] \hat{\sigma}^2(X_i, Z_i), \quad (1.32) \]

where \( M \) is the number of matches per unit, \( K_M(i) \) is the number of times unit \( i \) is used as a match, and \( \hat{\sigma}^2(X_i, Z_i) \) is the estimated conditional variance given by

\[ \hat{\sigma}^2(X_i, Z_i) = \frac{J}{J + 1} \left( Y_i - \frac{1}{J} \sum_{m=1}^{J} Y_{m(i)} \right)^2, \]

where \( J \) is a fixed number of similar units and \( l_m(i) \) is the \( m \)th closest unit to unit \( i \) among units with the same \( Z \) value [1].

A final strategy for estimating the variance of matching estimators is bootstrapping, which seeks to estimate the sampling distribution of the matching estimator. For each bootstrap sample, the propensity score is estimated, matches are made, and an estimate created. Though time intensive, bootstrapping seems appealing as a way to include the variance of the propensity score estimation and matching procedures. However, Abadie and Imbens show that bootstrapping does not lead to valid confidence intervals for nearest-neighbor matching with a fixed number of matches in general [2].

### 1.5.2 Stratification Estimators

Typically in practice, the variance of a stratification estimator is estimated by treating the within-propensity-score stratum effect estimates as independent estimates of the treatment effect, and then pooling the variance across strata [27]. Within a propensity score stratum, the treatment and control estimators are assumed to be
independent. The total variance is the sum of the variance of the two estimators. For example, take the stratified estimator $\hat{\Delta}_S$ from (1.19). A variance estimate is

$$v\hat{a}r(\hat{\Delta}_s) = \sum_{i=1}^{K} \left( \frac{n_k}{n} \right)^2 \hat{\sigma}_k^2$$

(1.33)

where

$$\hat{\sigma}_k^2 = n_{1k}^{-1} s_{1k}^2 + n_{0k}^{-1} s_{0k}^2,$$

and

$$s_{1k}^2 = n_{1k}^{-1} \sum_{i=1}^{n} I(\hat{e}_i \in Q_k)(z_i y_i - \bar{y}_{1k})^2$$

and

$$s_{0k}^2 = n_{0k}^{-1} \sum_{i=1}^{n} I(\hat{e}_i \in Q_k)((1 - z_i)Y_i - \bar{y}_{0k})^2$$

and

$$\bar{y}_{1k} = \sum_{i=1}^{n} I(\hat{e}_i \in Q_k)z_i y_i$$

and

$$\bar{y}_{0k} = \sum_{i=1}^{n} I(\hat{e}_i \in Q_k)(1 - z_i)y_i.$$  

If we are interested in estimating $\Delta_{ATT}$ instead, simply use

$$v\hat{a}r(\hat{\Delta}_{ts}) = \sum_{i=1}^{K} \left( \frac{n_{1k}}{n_{1}} \right)^2 \hat{\sigma}_k^2$$

(1.34)

For the model-assisted estimator $\hat{\Delta}_{sr}$, a variance estimator is

$$v\hat{a}r(\hat{\Delta}_{sr}) = \sum_{i=1}^{K} \left( \frac{n_k}{n} \right)^2 \hat{\sigma}_k^2,$$

(1.35)

but now $\hat{\sigma}_k^2$ is the estimate of the variance of the regression estimator in the $k$th stratum.

These variance estimators have been found to do well in many situations, though possibly slightly conservative in certain scenarios [27] [45]. Williamson et al. [45] investigate the impact of including the propensity score estimates in the variance estimator by treating the propensity score parameters, strata boundaries, fraction of the population in each stratum, and the stratified treatment effect as solutions to a set of estimating equations. They derive the large sample marginal variance of the stratified estimator through M-estimation theory, although they conclude that in
real-life situations this derivation is too complex and bootstrapping should be used instead.

1.5.3 IPW Estimators

The variance of IPW estimators can be estimated by viewing the estimates as solutions to a set of estimating equations and using the resulting empirical sandwich estimator of variance. Hernan and Robins [13] recommend fitting a generalized estimating equations model with an independent working covariance matrix to estimate the variance.

Define $W$ as the $n \times n$ weight matrix where $W_{ij} = 0$ $\forall$ $i \neq j$, and

$$W_{ii} = w_i = \left( \frac{1}{z_i w_i^* + (1 - z_i) w_0^*} \right) \frac{1}{z_i (\hat{e}_i) + (1 - z_i)(1 - \hat{e}_i)},$$

where $w_i^* = \sum_{i=1}^{n} z_i / \hat{e}_i$ and $w_0^* = \sum_{i=1}^{n} (1 - z_i) / (1 - \hat{e}_i)$. Also define $A' = (1, z)$, a $n \times 2$ matrix, and $\eta = (\mu_0, \Delta)'$, where $z$ is the vector of observed treatment exposures. Then the IPW estimator $\hat{\Delta}_{ipw2}$ (1.24) can be computed using the weighted least-squares estimator

$$\hat{\eta} = (\hat{\mu}_0, \hat{\Delta})' = (A'W A)^{-1} A'W y,$$

where $y$ is the vector of observed outcomes. The GEE empirical sandwich estimator for the variance of $\hat{\eta}$ is given by

$$\widehat{Var}(\hat{\eta}) = \left\{ (A'W A)^{-1}(A'W \hat{V} W A)(A'W A)^{-1} \right\},$$

where $\hat{V}$ is an empirical estimate of the covariance of $y$. The GEE model implies that we estimate $\hat{V}$ by the $n \times n$ diagonal matrix of squared residuals, $R$, where

$$R_{ii} = r_i^2 = (y_i - A_i' \hat{\eta})^2.$$
The GEE model variance estimate for $\hat{\Delta}_{ipw2}$ reduces to

$$\hat{\text{Var}}(\hat{\Delta}_{ipw2}) = \sum_{i=1}^{n} r_i^2 w_i^2.$$  

(1.36)

While this estimator of variance is routinely used, it is often far too conservative [45]. Lunceford and Davidian [27] include the propensity score model in the estimating equations, which makes the estimation more exact. The theory of M-estimation implies that for IPW estimators $\hat{\Delta}$, $n^{1/2}(\hat{\Delta} - \Delta)$ converges in distribution to a $N(0, \Sigma)$ variable. M-estimators are solutions to the vector equation $\sum_{i=1}^{n} \psi(Y_i, \theta) = 0$, where $Y_i, \ldots, Y_n$ are independent (but not necessarily identically distributed) variables, $\theta$ is a $p$-dimensional parameter, and $\psi$ is a known $(p \times 1)$-function that does not depend on $i$ or $n$ [41].

Under suitable regularity conditions $\hat{\theta}$ is asymptotically multivariate normal as $n \to \infty$, with mean $\theta_0$, the true value of $\theta$ and variance $(1/n) V(\theta_0)$, where

$$V(\theta_0) = A(\theta_0)^{-1} B(\theta_0) \left\{ A(\theta_0)^{-1} \right\}',$$

$$A(\theta_0) = E \left[ - \frac{\partial \psi(Y, \theta)}{\partial \theta} \bigg| \theta = \theta_0 \right],$$

and

$$B(\theta_0) = E \left[ \psi(Y, \theta_0) \psi(Y, \theta_0)' \right].$$

If we assume for the moment that the true propensity scores are known, or equivalently that the coefficients of true propensity score model are known, then for example $\hat{\Delta}_{ipw1}$,

$$\psi(Y_i, \hat{\theta}) = \frac{Z_i Y_i}{e_i} + \frac{(1 - Z_i) Y_i}{1 - e_i} - \hat{\Delta}.$$
It then follows that

\[ A(\theta_0) = -1, \text{ and} \]

\[
B(\theta_0) = E \left[ \left( \frac{ZY}{e} + \frac{(1-Z)Y}{1-e} - \Delta \right)^2 \right] \\
= E \left[ \left( \frac{ZY^z=1}{e} \right)^2 + \left( \frac{(1-Z)Y^z=0}{1-e} \right)^2 \right] + \hat{\Delta}^2 - 2\hat{\Delta} E \left[ \frac{ZY^z=1}{e_i} + \frac{(1-Z)Y^z=0}{1-e_i} \right] \\
= E \left[ \frac{Z(Y^z=1)^2}{e^2} + \frac{(1-Z)(Y^z=0)^2}{(1-e)^2} \right] + \hat{\Delta}^2 - 2\hat{\Delta}^2 \\
= E \left[ \frac{(Y^z=1)^2}{e} + \frac{(Y^z=0)^2}{1-e} \right] - \hat{\Delta}^2.
\]

This implies that the large sample variance of \( \hat{\Delta}_{ipw1} \) is

\[ \Sigma_{ipw1}^* = E \left[ \frac{(Y^z=1)^2}{e} + \frac{(Y^z=0)^2}{1-e} \right] - \hat{\Delta}^2. \]

Similarly the large-sample variance for \( \hat{\Delta}_{ipw2} \) and \( \hat{\Delta}_{ipw3} \) are

\[ \Sigma_{ipw2}^* = E \left[ \frac{(Y^z=1 - \mu_1)^2}{e} + \frac{(Y^z=0 - \mu_0)^2}{1-e} \right] \text{ and} \]

\[ \Sigma_{ipw3}^* = E \left[ \frac{(Y^z=1 - \mu_1)^2}{e} + \frac{(Y^z=0 - \mu_0)^2}{1-e} \right] + \eta_1 E \left( \frac{Y_1 - \mu_1}{e} \right) + \eta_0 E \left( \frac{Y_0 - \mu_0}{1-e} \right) + 2\eta_1 \eta_0, \]

where \( \mu_1 = E[Y^z=1], \mu_0 = E[Y^z=0], \eta_1 = -E\{Z(Y - \mu_1)/e^2\}/E\{(Z - e)^2/e^2\}, \) and \( \eta_0 = -E\{(1-Z)(Y - \mu_0)/(1-e)^2\}/E\{(Z - e)^2/(1-e)^2\}. \)

Now, instead assume that we estimate the propensity scores using the logistic regression model \( e(X, \beta) = \{1 + exp(-X'\beta)\}^{-1} \) and estimate \( \beta \) by maximum likelihood by solving

\[
\sum_{i=1}^n \psi(Z_i, X_i, \beta) = \sum_{i=1}^n \frac{Z_i - e(X_i, \beta)}{e(X_i, \beta)(1 - e(X_i, \beta))} \frac{\partial}{\partial \beta} e(X_i, \beta) = 0. \tag{1.37}
\]
Then for example $\hat{\Delta}_{ipw1}$ we have that,

$$
\psi(Y_i, \hat{\theta}) = \left( \frac{Z_i Y_i}{e_i} + \frac{(1-Z_i)Y_i}{1-e_i} - \hat{\Delta} \right)
$$

The large sample variances of $\hat{\Delta}_{ipw1}, \hat{\Delta}_{ipw2}, \hat{\Delta}_{ipw3}, \hat{\Delta}_{dr}$ then become

$$
\Sigma_{ipw1} = \Sigma_{ipw1}^* - \mathbf{H}_{\beta,1} \mathbf{E}_{\beta}^{-1} \mathbf{H}_{\beta,1}, \quad \mathbf{H}_{\beta,1} = E \left[ \left( \frac{Y_{z=1}}{e} + \frac{Y_{z=0}}{1-e} \right) e_{\beta} \right],
$$

(1.38)

$$
\Sigma_{ipw2} = \Sigma_{ipw2}^* - \mathbf{H}_{\beta,2} \mathbf{E}_{\beta}^{-1} \mathbf{H}_{\beta,2}, \quad \mathbf{H}_{\beta,2} = E \left[ \left( \frac{Y_{z=1} - \mu_1}{e} + \frac{Y_{z=0} - \mu_0}{1-e} \right) e_{\beta} \right],
$$

(1.39)

$$
\Sigma_{ipw3} = \Sigma_{ipw3}^* - \mathbf{H}_{\beta,3} \mathbf{E}_{\beta}^{-1} \mathbf{H}_{\beta,3}, \quad \mathbf{H}_{\beta,3} = E \left[ \left( \frac{Y_{z=1} - \mu_1 + \eta_1}{e} + \frac{Y_{z=0} - \mu_0 + \eta_0}{1-e} \right) e_{\beta} \right],
$$

(1.40)

$$
\Sigma_{dr} = \Sigma_{ipw2}^* - E \left[ \sqrt{\frac{1-e}{e}} (E[Y_{z=1}|X] - \mu_1) + \sqrt{\frac{e}{1-e}} (E[Y_{z=0}|X] - \mu_0) \right]^2,
$$

(1.41)

where $e_{\beta} = \frac{\partial}{\partial \beta} e(X, \beta)$, and $E_{\beta\beta} = E[\frac{e e'_{\beta}}{e(1-e)}]$. Interestingly, estimating $\beta$ leads to a smaller large sample variance than using the true value [27].

We can either estimate the components in (1.38) - (1.41) by the observed data, or use the empirical estimators of $A(\theta_0)$ and $B(\theta_0)$

$$
A_n(Y, \hat{\theta}) = \frac{1}{n} \sum_{i=1}^{n} \left[ -\frac{\partial \psi(Y_i, \theta)}{\partial \theta'} \right]_{\theta=\hat{\theta}}, \quad \text{and}
$$

$$
B_n(Y, \hat{\theta}) = \frac{1}{n} \sum_{i=1}^{n} \left[ \psi(Y_i, \hat{\theta}) \psi(Y_i, \hat{\theta})' \right].
$$
Using these empirical estimates yields the empirical sandwich variance estimator. This method is preferred because variance estimates are more stable in practice. Specifically, the empirical sandwich estimators for the variance of $\hat{\Delta}_{ipw1}$, $\hat{\Delta}_{ipw2}$, $\hat{\Delta}_{ipw3}$, $\hat{\Delta}_{dr}$ are given by

$$n^{-2} \sum_{i=1}^{n} \hat{I}_{i}^2, \quad \text{where}$$ (1.42)

$$\hat{I}_{ipw1,i} = \frac{Z_{i}Y_{i}}{\hat{e}_{i}} - \frac{(1 - Z_{i})Y_{i}}{1 - \hat{e}_{i}} - \hat{\Delta}_{ipw1} - (Z_{i} - \hat{e}_{i})\hat{H}_{\beta,1}^{\prime} \hat{E}_{\beta\beta}^{-1} \mathbf{X}_{i},$$ (1.43)

$$\hat{I}_{ipw2,i} = \frac{Z_{i}(Y_{i} - \hat{\mu}_{1,ipw2})}{\hat{e}_{i}} - \frac{(1 - Z_{i})(Y_{i} - \hat{\mu}_{0,ipw2})}{1 - \hat{e}_{i}} - (Z_{i} - \hat{e}_{i})\hat{H}_{\beta,2}^{\prime} \hat{E}_{\beta\beta}^{-1} \mathbf{X}_{i},$$ (1.44)

$$\hat{I}_{ipw3,i} = \frac{Z_{i}(Y_{i} - \hat{\mu}_{1,ipw3} + \hat{\eta}_{1}(Z_{i} - \hat{e}_{i}))}{\hat{e}_{i}} - \frac{(1 - Z_{i})(Y_{i} - \hat{\mu}_{0,ipw3} + \hat{\eta}_{0}(Z_{i} - \hat{e}_{i}))}{1 - \hat{e}_{i}}$$
$$- (Z_{i} - \hat{e}_{i})\hat{H}_{\beta,3}^{\prime} \hat{E}_{\beta\beta}^{-1} \mathbf{X}_{i},$$ (1.45)

$$\hat{I}_{dr,i} = \frac{Z_{i}Y_{i} - m_{1}(\mathbf{X}_{i}, \hat{\alpha}_{1})(Z_{i} - \hat{e}_{i})}{\hat{e}_{i}} - \frac{(1 - Z_{i})Y_{i} + m_{0}(\mathbf{X}_{i}, \hat{\alpha}_{0})(Z_{i} - \hat{e}_{i})}{1 - \hat{e}_{i}} - \hat{\Delta}_{dr},$$ (1.46)

$$\hat{E}_{\beta\beta}^{-1} = n^{-1} \sum_{i=1}^{n} \hat{e}_{i}(1 - \hat{e}_{i}) \mathbf{X}_{i} \mathbf{X}_{i}^{\prime},$$

$$\hat{H}_{\beta,1} = n^{-1} \sum_{i=1}^{n} \left\{ \frac{Z_{i}Y_{i}(1 - \hat{e}_{i})}{\hat{e}_{i}} + \frac{(1 - Z_{i})Y_{i}\hat{e}_{i}}{1 - \hat{e}_{i}} \right\} \mathbf{X}_{i},$$

$$\hat{H}_{\beta,2} = n^{-1} \sum_{i=1}^{n} \left\{ \frac{Z_{i}(Y_{i} - \hat{\mu}_{1,ipw2})(1 - \hat{e}_{i})}{\hat{e}_{i}} + \frac{(1 - Z_{i})(Y_{i} - \hat{\mu}_{0,ipw2})\hat{e}_{i}}{1 - \hat{e}_{i}} \right\} \mathbf{X}_{i},$$
\[
\hat{H}_{\beta,3} = n^{-1} \sum_{i=1}^{n} \left\{ \frac{Z_i (Y_i - \hat{\mu}_{1,ipw3} + \hat{\eta}_1) (1 - \hat{e}_i)}{\hat{e}_i} + \frac{(1 - Z_i) (Y_i - \hat{\mu}_{0,ipw3} + \hat{\eta}_0) \hat{e}_i}{1 - \hat{e}_i} \right\} X_i,
\]

\[
\hat{\eta}_1 = -\frac{\sum_{i=1}^{n} \{Z_i (Y_i - \hat{\mu}_{1,ipw3}) / \hat{e}_i^2 \}}{\sum_{i=1}^{n} \{(Z_i - \hat{e}_i) / \hat{e}_i \}^2}, \quad \text{and} \quad \hat{\eta}_0 = -\frac{\sum_{i=1}^{n} \{(1 - Z_i) (Y_i - \hat{\mu}_{0,ipw3}) / (1 - \hat{e}_i)^2 \}}{\sum_{i=1}^{n} \{(Z_i - \hat{e}_i) / (1 - \hat{e}_i) \}^2}.
\]

The corresponding estimators for the variance of \(\hat{\Delta}_{ipw1}^t, \hat{\Delta}_{ipw2}^t, \) and \(\hat{\Delta}_{ipw3}^t\) are given by

\[
\hat{\Delta}_{ipw1}^t = n^{-1} \sum_{i=1}^{n} \hat{I}_{ipw1,i}^2, \quad \text{where} \quad \hat{I}_{ipw1,i} = Z_i Y_i - \frac{(1 - Z_i) Y_i \hat{e}_i}{1 - \hat{e}_i} - \hat{\Delta}_{ipw1}^t - (Z_i - \hat{e}_i) \hat{H}_{\beta,1}^{-1} \hat{E}_{\beta\beta}^{-1} X_i,
\]

\[
\hat{I}_{ipw2,i} = Z_i (Y_i - \hat{\mu}_{1,ipw2}^t) - \frac{(1 - Z_i) (Y_i - \hat{\mu}_{0,ipw2}) \hat{e}_i}{1 - \hat{e}_i} - (Z_i - \hat{e}_i) \hat{H}_{\beta,2}^{-1} \hat{E}_{\beta\beta}^{-1} X_i,
\]

\[
\hat{I}_{ipw3,i} = Z_i (Y_i - \hat{\mu}_{1,ipw3}^t) - \frac{(1 - Z_i) (Y_i - \hat{\mu}_{0,ipw3}) \hat{e}_i - \hat{\eta}_0 \hat{e}_i (Z_i - \hat{e}_i)}{1 - \hat{e}_i} - (Z_i - \hat{e}_i) \hat{H}_{\beta,3}^{-1} \hat{E}_{\beta\beta}^{-1} X_i
\]

\[
\hat{E}_{\beta\beta}^{-1} = n^{-1} \sum_{i=1}^{n} \hat{e}_i (1 - \hat{e}_i) X_i X_i',
\]

\[
\hat{H}_{\beta,1} = n^{-1} \sum_{i=1}^{n} \left\{ \frac{(1 - Z_i) Y_i \hat{e}_i}{1 - \hat{e}_i} \right\} X_i,
\]

30
\[ \hat{H}_{\beta,2} = n^{-1} \sum_{i=1}^{n} \left\{ \frac{(1-Z_i)(Y_i - \hat{\mu}_{0,\text{ipw}2})\hat{e}_i}{1 - \hat{e}_i} \right\} X_i, \]

\[ \hat{H}_{\beta,3} = n^{-1} \sum_{i=1}^{n} \left\{ \frac{(1-Z_i)(Y_i - \hat{\mu}_{0,\text{ipw}3} + \hat{\eta}_0)\hat{e}_i}{1 - \hat{e}_i} + \hat{\eta}_0\hat{e}_i(1 - \hat{e}_i) \right\} X_i, \]

\[ \hat{\eta}_0 = -\frac{\sum_{i=1}^{n} \{ (1-Z_i)(Y_i - \hat{\mu}_{0,\text{ipw}3})\hat{e}_i/(1 - \hat{e}_i)^2 \}}{\sum_{i=1}^{n} \{ (Z_i - \hat{e}_i)\hat{e}_i/(1 - \hat{e}_i) \}^2}. \]

We note that the variance estimators for \( \hat{\Delta}_{\text{ipw}1}^t, \hat{\Delta}_{\text{ipw}2}^t, \) and \( \hat{\Delta}_{\text{ipw}3}^t \) above have not been published in the statistics literature at this point to our knowledge.

### 1.6 Complex Survey Sampling

In complex survey sampling, researchers specify a sampling design with possibly multiple levels of stratification and clustering. In addition, specific groups may be oversampled so that more precise inference can be made about them. Large sample surveys seek to make inference about a population or subpopulations. In order for the sample to be representative of the population, we must account for the sample design used to generate the survey. Typically this is done by assigning weights to observations based on the inverse probability of being included in the sample. The weights and the sampling design are used along with the observations to make inferences about the population parameters through inverse sampling probability estimators such as the Horvitz-Thompson estimator [19]. Variance estimates are often made using linearization methods or by replication weights [26].

In addition to weights created by the inverse probability of selection, non-response and post-stratification adjustments are often made in an effort to align the distributions of the sample to known population totals. In total, the final weight assigned to a sample observation takes into account the probability of selection, non-response, and
post-stratification adjustments. Inference that ignores the survey design may be subject to bias and incorrect variance estimates when attempting to estimate population parameters.

1.7 Literature Review of Propensity Score Methods with Complex Survey Data

Research on propensity score methods for complex survey data is somewhat limited, with only a few authors offering work on the subject. In practice there is not a consensus on when and how to use the survey design information when attempting to estimate the population average treatment effect. DuGoff et al. [9] note that many studies using complex survey data in health services research do not incorporate the survey weights into their analysis. They find that those who do so vary in their interpretation of the resulting estimate, suggesting “some possible misunderstandings in how to appropriately apply and interpret results from propensity score methods with complex survey data.” Methods resembling propensity score stratification, weighting, and matching have been proposed in the complex survey data context; however, little justification for these methods has been shown through theory or simulation.

One of the first steps in any propensity score analysis is to estimate the propensity score for the observed units. This is the first place where the survey design could possibly be included. In two papers, Zanutto et al. [48] and Zanutto [49] argue that the sampling weights should not be used in the propensity score analysis because “the estimated propensity scores are used only to form subclasses with similar background covariates in the sample data and not to make inferences about the population-level propensity score model, it is not necessary to use survey-weighted estimation for the propensity score model”. Zanutto and others go on to say that survey design variables should be included in the model, and that survey weights may be used as a covariate.
if the cluster or stratum identifiers are not released in the public data. Dugoff et al. [9] agree with Zanutto in their work, estimating the propensity score without the survey design. The only author who does use a survey-weighted model to estimate the propensity score is Rubin [39], but he does not offer any rationale for the decision.

Zanutto proposes stratified estimators that adjust for complex survey data [49] and compares them to using a survey-weighted regression analysis. Her proposed estimator which parallels (1.20) is given by:

$$\hat{\Delta}_{ws} = \frac{5}{\sum_{k=1}^{K} \sum_{i \in S_{Fk}} w_i} \left( \sum_{i \in S_{Mk}} w_i y_i - \sum_{i \in S_{Fk}} w_i y_i \right)^2,$$  \hspace{1cm} (1.51)$$

where $w_i$ represents the survey weight of unit $i$, $S_{Fk}$ and $S_{Mk}$ denote the set of treatment units and control units in stratum $k$ (males and females in the paper). The response for unit $i$ is written as $y_i$, and there are $K = 5$ propensity score strata. Here the estimator is weighted to the treated population so that it estimates $\Delta_{PATT}$.

Rubin [39] writes about a similar idea, but is not as explicit.

The proposed variance estimate for $\hat{\Delta}_{ws}$ is given by

$$\hat{\text{var}}(\hat{\Delta}_{ws}) = \sum_{k=1}^{K} \left( \frac{\sum_{i \in S_{Fk}} w_i}{\sum_{k=1}^{K} \sum_{i \in S_{Fk}} w_i} \right)^2 \left( s_{Mk}^2 + s_{Fk}^2 \right),$$  \hspace{1cm} (1.52)$$

where

$$s_{Mk}^2 = \frac{n}{n-1} \sum_{i=1}^{n} \left( z_{ik} - \frac{1}{n} \sum_{j=1}^{n} z_{jk} \right)^2,$$

and

$$z_{ik} = \begin{cases} \frac{w_i}{\sum_{i \in S_{Mk}} w_i} \left( y_i - \frac{\sum_{i \in S_{Mk}} w_i y_i}{\sum_{i \in S_{Mk}} w_i} \right) & i \in S_{Mk} \\ 0 & i \notin S_{Mk} \end{cases}.$$
Note that estimates of the relative size of propensity score strata (estimated by the sum of the weights) are treated as fixed in the variance estimate, and secondly the survey design features such as a sampling stratum or cluster are not used in the formula, only the sampling weights.

Zanutto also proposes a regression-adjusted stratified estimator comparable to equation (1.21). Within each propensity score stratum, a survey-weighted regression model is estimated in an effort to correct for remaining imbalances. The estimator is given by

$$\hat{\Delta}_{w_{sr}} = \sum_{k=1}^{5} \left( \frac{\sum_{i \in S_{Fk}} w_i}{\sum_{k=1}^{5} \sum_{i \in S_{Fk}} w_i} \right) \hat{\beta}_{w}^{k}, \quad (1.53)$$

where $\hat{\beta}_{w}^{k}$ is the coefficient of the indicator variable for treatment in the regression model for stratum $k$. An estimator for the variance of $\hat{\Delta}_{w_{sr}}$ is given by

$$\hat{\Delta}_{w_{sr}} = \sum_{k=1}^{5} \left( \frac{\sum_{i \in S_{Fk}} w_i}{\sum_{k=1}^{5} \sum_{i \in S_{Fk}} w_i} \right)^2 \hat{V}(\hat{\beta}_{w}^{k}) \quad (1.54)$$

Several authors consider survey-based adjustments to inverse probability weighting estimators [39][48][9]. The general idea is to take the weight used in the typical IPW estimator and multiply it by the survey weight to form a new weight. As with all weighting techniques, researchers are wary of possibly large weights created by the multiplication of the two weights.

No author proposes a matching-based method that accommodates the sampling design and utilizes the matched pairs for estimation. Rubin [39] uses matching in his analysis to create better balance, but then uses a stratification estimator. Dugoff et al. [9] matches on the propensity score, but then estimates the treatment effect using a survey-weighted linear regression model based on the data from the sub-sample of matched pairs and their accompanying sampling weight.
Dugoff et al. [9] use survey-weighted regression for all types of proposed propensity score methods: matching, stratification, and weighting. For stratification methods they perform a survey-weighted regression inside each propensity score stratum. In place of a weighting estimator they estimate a survey-weighted regression model with the product of the propensity-score weights and sampling weights. All of their proposed methods are assisted by regression modeling.

The only simulation study that attempts to justify use of survey-adjusted propensity score estimators is by Dugoff et al. [9], but the setup is very simple, making it difficult to lend much justification to the proposed methods. Only a single covariate \(X\) is used as a confounder, which implies that the propensity score is a 1-1 function of \(X\). Stratifying or matching on the propensity score is actually the same as stratifying or matching on \(X\). Also note that, as we have discussed in the previous paragraph, all of the the methods utilize regression modeling in some way. The simulation results are impressive in terms of absolute bias and coverage compared to the non-survey adjusted methods, but we argue that more complicated situations should be tested with estimators that are not aided by a regression model.

While the simulation in [9] is very simplistic, it does highlight an important fact that seems to be overlooked in the literature pertaining to propensity score methods in complex survey data. The simulation uses an interaction effect so that the treatment effect is not homogeneous across the population. In fact, this interaction effect was very strong. This attributed greatly to the bias in the non-survey adjusted estimates. Without this interaction effect, the non-survey adjusted estimates should have small bias or be unbiased.
Variance estimation is almost completely unstudied in this context. While some variance estimators are proposed [49] for survey-adjusted propensity score methods, there is no investigation into their performance.
Chapter 2: General Framework, Weighting and Stratification Estimators

In this chapter we provide a theoretical basis for methods that incorporate survey design features into a propensity score analysis. First we propose a framework from which to think about the problem in terms of sampling in two steps. Next, we derive stratification and weighting estimators, and we discuss their properties. We then discuss propensity score estimation and propose variance estimators for each of the stratification and weighting estimators. To begin with we will focus on estimators for \( \Delta_{PATE} \) and then later describe estimators of \( \Delta_{PATT} \). Lastly we create a simulation to investigate the properties of our estimators and compare them against others.

2.1 A General Framework

Consider a potential outcomes super-population in which a pair of potential outcomes \((Y_{i}^{z=1}, Y_{i}^{z=0})\) exist for \(i = 1, \ldots, N\) units. For each pair \(i\), there exists a fixed set of covariates \(X_{i}\). Furthermore, let \(X\) be subdivided into three sets of covariates \(\{X\} = \{X^{Z}, X^{S}, X^{ZS}\}\), which affect the probability of treatment selection only, sample selection only, and treatment selection and sample selection both, respectively.

Now consider a population sampled from the super-population of \(N\) units, which consists of exactly one of the two potential outcomes and associated covariates for each unit in the super-population. Assume there exists a true population propensity score,
\[ e_i = P(Z_i = 1), \text{ that is fixed for all units. } Z_i = 1 \text{ is equivalent to saying that } Y_{iz=1} \text{ is sampled instead of } Y_{iz=0} \text{ and vice versa for } Z_i = 0. \] Let \( e_i \) be solely determined by the covariates \( X_i \), and \( e(X) = P(Z = 1|X^Z, X^{ZS}) \) so that units with the same covariate pattern share the same propensity score. Also assume that \( 0 < e_i < 1 \forall i = 1, \ldots, N \).

Now consider a population of \( N \) observations that has been sampled from the super-population based on the propensity scores, and suppose we take a random probability sample of size \( n \) of the population based on some complex sample design with the probability of selection known for each unit in the population. Define \( S_i \) as the indicator that unit \( i \) in the population has been sampled, and let the probability of sampling be determined by \( X \) and possibly \( Z \), so that \( P(S = 1|X^S, X^{ZS}, Z) \) is known. Define \( f(X, Z) = P(S = 1|X^S, X^{ZS}, Z) \), and let \( \sum_{i=1}^{N} f(X_i, Z_i) = n \). Let \( f_i \) indicate the sampling probability of unit \( i \) and assume \( 0 < f_i < 1 \forall i = 1, \ldots, N \). If the probability of sampling is determined by \( Z \), then note that \( f_i \) can change across different first stage samples, making it a random variable.

It is possible for a researcher to choose a treatment variable \( Z \) that directly determines the probability of sampling, but for the most part we believe that this scenario is less common. It will be more common that \( S \) and \( Z \) will share common predictors, but they will be conditionally independent. In the following proofs we will assume that \( S \perp Z|X \), which implies that \( f_i \) is fixed for \( i \in 1, \ldots, N \). The results or steps of the proofs still hold even if this assumption is not true, and the assumption makes the notation of the proof much simpler. We illustrate this assumption through a short example. Say we wish to estimate the effect of having Medicaid health insurance vs. being uninsured on health outcomes in a survey that over-represents those with Medicaid by over-sampling certain geographical regions and race-ethnicities. The probability of treatment (being covered by Medicaid) and being sampled are
not marginally independent, but they will be independent conditional on geographic region and race-ethnicity.

Multiplying the probability of treatment and that of being sampled together allows us to define the probability that a particular potential outcome is included in our sample by

$$P(S = 1, Z = z|X) = P(Z = z|X)P(S = 1, |Z = z, X)$$ (2.1)

When we refer to the variables belonging to unit $i$ in the population, that is $(Y_i, Z_i, S_i, X_i)$, $Y_i, Z_i$ and $S_i$ are all random random variables, while $X_i$ are fixed characteristics of that unit. This is similar to a sampling framework, with the only difference being that the variable $Y_i$ is random because it takes on different values depending on the value of $Z_i$. This implies that for example $E[Z_i = 1] = E[Z_i = 1|X_i]$ because the covariate pattern $X$ is fixed for unit $i$. $Y_i$ and $Z_i$ do become fixed after the first stage of sampling, while $S_i$ would still be random at that point.

In this perspective, the propensity score links the sample, the study population, and the super-population, which allows us to estimate the average treatment effect in ways that will be reflective of the population distribution. See Figure 2.1 for a graphical representation of the framework. In comparison, while a conventional propensity score analysis may consider the potential outcomes for the sample, the potential outcomes for the population are ignored. See Figure 2.2 for a graphical representation of a typical potential outcomes framework. If the sample is indeed a simple random sample from the population, then ignoring the sampling design should lead to an unbiased estimate of the population average treatment effect, but otherwise estimation of the treatment effect will likely be biased.
2.1.1 Sampling Assumptions

In general, we are thinking about sampling designs that use a stratification in the first stage rather than a first stage cluster design. A design which includes a first stage cluster creates a large dependence between sampled observations complicating matters. Later we consider this issue more throughly, but for the moment assume that any clustering in the design is in a later stage.

2.2 Stratification Estimators

Stratified estimation follows naturally from our sampling framework. We create strata under the assumption that a coarse matching of the observations by the propensity score eliminates bias due to treatment imbalance. As a result, after stratification we have strongly ignorable treatment assignment [36] within each stratum. Weighting the within-stratum estimates by the proportion of the population each represents gives an estimate of the average population treatment effect. Define $\Delta_{s,svy}$ as

$$
\Delta_{s,svy} = \sum_{k=1}^{K} \frac{N_k}{N} \left[ \frac{1}{N_k\tilde{e}_k} \sum_{i=1}^{N} Z_i S_i I(e_i \in Q_k) \frac{Y_i}{f_i} - \frac{1}{N_k(1-\tilde{e}_k)} \sum_{i=1}^{N} (1-Z_i) S_i I(e_i \in Q_k) \frac{Y_i}{f_i} \right],
$$

(2.2)

where $K$ strata are formed by grouping the observations by $e$, the true propensity score. The propensity score strata are referred to as $Q_k$, $k = 1, \ldots, K$, with constant propensity score $\tilde{e}_k$ within each stratum. This is a similar estimator to (1.51).

Let $Z = [Z_1, \ldots, Z_N]'$ be the random vector of length $N$ indicating which potential outcomes are chosen in the first stage of sampling. The probability that a certain sample of potential outcomes is chosen in the first stage of sampling is given by
\[
P(Z = z) = \prod_{i=1}^{N} \{z_i e_i + (1 - z_i)(1 - e_i)\}.
\]

Similarly let \( S = [S_1, \ldots, S_N]' \) denote the population units sampled in the second stage of sampling. Then \((Z, S)\) specifies a specific sample of potential outcomes and units.

The following statements summarize the assumptions we make in our estimation of \( \Delta_{PATE} \):

(A.1) The total population size \( N \) is known.

(A.2) The true population propensity score, \( 0 < e_i < 1 \), for each unit \( i = 1, \ldots, N \) is known.

(A.3) \( (Y^{z=1}, Y^{z=0}) \perp \perp Z|e \).

(A.4) The true probability of sample selection \( 0 < f_i < 1 \), for each unit \( i = 1, \ldots, N \) is known.

(A.5) There exist \( K \) unique propensity scores \( \tilde{e}_1, \ldots \tilde{e}_K \), such that when divided into \( K \) propensity score strata all units in a stratum have equal propensity score \( \tilde{e}_k \).

(A.6) The size of the propensity score strata \( N_k, k = 1, \ldots, K \) are known.

**Proposition 1.** Assuming (A.1) through (A.6), \( \Delta_{s, svy} \) is an unbiased estimator of \( \Delta_{PATE} \).

**Proof.**

For each individual \( i \) we can write \( Y_i \) as

\[
Y_i = Z_i S_i Y_i^{z=1} + Z_i (1 - S_i) Y_i^{z=1} + (1 - Z_i) S_i Y_i^{z=0} + (1 - Z_i)(1 - S_i) Y_i^{z=0},
\]
which implies that \( Z_i S_i Y_i = Z_i S_i Y_i^{z=1} \) and \( (1 - Z_i) S_i Y_i = (1 - Z_i) S_i Y_i^{z=0} \) because 
\[ Z_i Z_i = Z_i, \quad (1 - Z_i)(1 - Z_i) = (1 - Z_i), \quad Z_i(1 - Z_i) = 0, \quad S_i S_i = S_i, \quad \text{and} \quad S_i(1 - S_i) = 0. \]

We use these facts below. Taking the expectation of \( \frac{1}{N_k} \sum_{i=1}^{N} \frac{Z_i S_i I(e_i \in Q_k) Y_i}{f_i} \) with respect to all possible samples \((Z, S)\), we show that it is equal to \( E[Y^{z=1}|e = \tilde{e}_k] = \frac{1}{N_k} \sum_{i=1}^{N} I(e_i = \tilde{e}_k) Y_i^{z=1}: \)

\[
E \left[ \frac{1}{N_k} \sum_{i=1}^{N} \frac{Z_i S_i I(e_i \in Q_k) Y_i}{f_i} \right] = \frac{1}{N_k} \sum_{i=1}^{N} E \left[ \frac{Z_i S_i I(e_i = \tilde{e}_k) Y_i^{z=1}}{f_i} \right] \\
= \frac{1}{N_k} \sum_{i=1}^{N} E \left[ Z_i S_i I(e_i = \tilde{e}_k) Y_i^{z=1} \right] \\
= \frac{1}{N_k} \sum_{i=1}^{N} E \left[ Z_i I(e_i = \tilde{e}_k) Y_i^{z=1} \right] \\
= \frac{1}{N_k} \sum_{i=1}^{N} E \left[ Z_i I(e_i = \tilde{e}_k) Y_i^{z=1} \right] \\
= \frac{1}{N_k} \sum_{i=1}^{N} E \left[ I(e_i = \tilde{e}_k) Y_i^{z=1} E[Z_i|X_i] \right] \\
= \frac{1}{N_k} \sum_{i=1}^{N} I(e_i = \tilde{e}_k) Y_i^{z=1} \tilde{e}_k \\
= \frac{1}{N_k} \sum_{i=1}^{N} I(e_i = \tilde{e}_k) Y_i^{z=1} \\
= E[Y^{z=1}|e = \tilde{e}_k]. \tag{2.3}
\]

Note that above \( E[Z_i I(e_i = \tilde{e}_k) Y_i^{z=1}] = I(e_i = \tilde{e}_k) Y_i^{z=1} E[Z_i|X_i] \) because \( I(e_i = \tilde{e}_k) \) and \( Y_i^{z=1} \) are fixed for unit \( i \) and \( E[Z_i] = E[Z_i|X_i] \) because \( X_i \) is a fixed characteristic of unit \( i \).

Similarly
\[
E \left[ \frac{1}{N_k(1 - \tilde{e}_k)} \sum_{i=1}^{N} \frac{(1 - Z_i)S_iI(e_i \in Q_k)Y_i}{f_i} \right]
\]
\[
= \frac{1}{N_k(1 - \tilde{e}_k)} \sum_{i=1}^{N} E \left[ (1 - Z_i)I(e_i = \tilde{e}_k) Y_{i}^{z=0} \right]
\]
\[
= E[Y_{i}^{z=0}|e = \tilde{e}_k]. \tag{2.4}
\]

Then taking (2.3) and (2.4) together we have that
\[
E \left[ \frac{1}{N_k(1 - \tilde{e}_k)} \sum_{i=1}^{N} \frac{Z_iS_iI(e_i \in Q_k)Y_i}{f_i} - \frac{1}{N_k(1 - \tilde{e}_k)} \sum_{i=1}^{N} \frac{(1 - Z_i)S_iI(e_i \in Q_k)Y_i}{f_i} \right]
\]
\[
= E[Y_{i}^{z=1} - Y_{i}^{z=0}|e = \tilde{e}_k].
\]

Note that \(N_k/N = P(e = \tilde{e}_k)\). Then averaging with respect to the distribution of \(e\) yields
\[
\sum_{k=1}^{K} \frac{N_k}{N} E \left[ \frac{1}{N_k(1 - \tilde{e}_k)} \sum_{i=1}^{N} \frac{Z_iS_iI(e_i \in Q_k)Y_i}{f_i} - \frac{1}{N_k(1 - \tilde{e}_k)} \sum_{i=1}^{N} \frac{(1 - Z_i)S_iI(e_i \in Q_k)Y_i}{f_i} \right]
\]
\[
= \sum_{k=1}^{K} P(e = \tilde{e}_k) E \left[ \frac{1}{N_k(1 - \tilde{e}_k)} \sum_{i=1}^{N} \frac{Z_iS_iI(e_i \in Q_k)Y_i}{f_i} - \frac{1}{N_k(1 - \tilde{e}_k)} \sum_{i=1}^{N} \frac{(1 - Z_i)S_iI(e_i \in Q_k)Y_i}{f_i} \right]
\]
\[
= E \left[ E[Y_{i}^{z=1} - Y_{i}^{z=0}|e = \tilde{e}_k] \right]
\]
\[
= E[Y_{i}^{z=1} - Y_{i}^{z=0}]
\]
\[
= \Delta_{PATE}.
\]

We note that in the above proof we implicitly assume that in all possible samples, we sample at least one treated and one control observation from each of the \(K\) propensity score strata. This issue also arises in post stratification literature [18].
In practice, \( N_k, \tilde{e}_k, k = 1, \ldots K \) are unknown as well as the propensity scores. If the true propensity scores are known, then

\[
E \left[ \sum_{i=1}^{N} \frac{Z_i S_i I(e_i \in Q_k)}{f_i} \right] = \sum_{i=1}^{N} E \left[ \frac{Z_i S_i I(e_i = \tilde{e}_k)}{f_i} \right] Z, X_i \\
= \sum_{i=1}^{N} E \left[ \frac{Z_i I(e_i = \tilde{e}_k)}{f_i} \right] E \left[ S_i | Z, X_i \right] \\
= \sum_{i=1}^{N} I(e_i = \tilde{e}_k) E[Z_i | X_i] \\
= N_k \tilde{e}_k.
\]

Similarly

\[
E \left[ \sum_{i=1}^{N} \left( 1 - Z_i \right) S_i I(e_i \in Q_k) \right] = N_k (1 - \tilde{e}_k), \text{ and}
\]

\[
E \left[ \sum_{i=1}^{N} \frac{(1 - Z_i) S_i I(e_i \in Q_k) + Z_i S_i I(e_i \in Q_k)}{f_i} \right] = E \left[ \sum_{i=1}^{N} S_i I(e_i \in Q_k) \right] = N_k.
\]

Using these as estimators of our unknown quantities and substituting estimated propensity scores for the true propensity scores leads to the estimator \( \hat{\Delta}_{s.svy} \), which is given by

\[
\hat{\Delta}_{s.svy} = \sum_{k=1}^{K} \frac{\hat{N}_{k1} + \hat{N}_{k0}}{N} \left[ \frac{1}{\hat{N}_{k1}} \sum_{i \in S} \frac{Z_i I(\hat{e}_i \in Q_k)Y_i}{f_i} - \frac{1}{\hat{N}_{k0}} \sum_{i \in S} \frac{(1 - Z_i) I(\hat{e}_i \in Q_k)Y_i}{f_i} \right],
\]

(2.5)

where

\[
\hat{N}_{k1} = \sum_{i \in S} \frac{Z_i I(e_i \in Q_k)}{f_i} \text{ and } \hat{N}_{k0} = \sum_{i \in S} \frac{(1 - Z_i) I(e_i \in Q_k)}{f_i}.
\]

The estimator \( \hat{\Delta}_{s.svy} \) is not guaranteed to be unbiased, even in the presence of strongly ignorable treatment assignment in each stratum. While \( \hat{N}_{k1} \) and \( \hat{N}_{k0} \) are unbiased estimators of \( N_k \tilde{e}_k \) and \( N_k (1 - \tilde{e}_k) \) under certain conditions,
are ratio estimators, and therefore typically not unbiased [26] (see 4.1.2). Ratio estimators have a bias on the order of $1/n$ in general ($1/n_{k_0}$ or $1/n_{k_1}$ in our case) [8].

So then, as long as the sample size is relatively large for both treatment and control groups in each stratum $k$, we expect the bias caused by using a ratio estimator to be small.

We recommend following the advice of Myers and Thomas [30] to determine the number of propensity score strata $K$. They recommend creating strata by quantiles of the propensity score and increasing the number of strata until the estimate remains relatively constant or until the estimated variance greatly increases. For a binary outcome with estimated proportion near zero or one, the number of strata will likely need to be smaller than if the outcome is continuous. This is so we can guarantee a minimum number of both outcome possibilities in a propensity score stratum.

### 2.3 Weighting Estimators

Next we consider weighting estimators based on the inverse probability of treatment similar to (1.22) and (1.24), but with adjustments for the survey design. As others have proposed [39] [48] [9], we consider weights defined by the product of the propensity-based weight and the survey weight. Define $\Delta_{ipw1.svy}$ as

$$\Delta_{ipw1.svy} = \frac{1}{N} \sum_{i=1}^{N} \frac{Z_i S_i Y_i}{f_i e_i} - \frac{1}{N} \sum_{i=1}^{N} \frac{(1 - Z_i) S_i Y_i}{f_i (1 - e_i)}. \quad (2.6)$$

**Proposition 2.** Under assumptions (A.1) through (A.4), $\Delta_{ipw.svy}$ is an unbiased estimator of $\Delta_{PATE}$. 

45
Proof.

\[
E \left[ \frac{1}{N} \sum_{i=1}^{N} \frac{Z_i S_i Y_i}{\hat{f}_i e_i} \right] = \frac{1}{N} \sum_{i=1}^{N} E \left[ \frac{Z_i S_i Y_i^z=1}{\hat{f}_i e_i} \right] = \frac{1}{N} \sum_{i=1}^{N} E \left[ \frac{Z_i Y_i^z=1}{\hat{f}_i e_i} \right] = \frac{1}{N} \sum_{i=1}^{N} E \left[ \frac{Z_i Y_i^z=1}{e_i} \right] = \frac{1}{N} \sum_{i=1}^{N} \frac{Y_i^z=1}{e_i} E[Z_i | X_i] = \frac{1}{N} \sum_{i=1}^{N} Y_i^z=1 = E[Y^z=1].
\]

Similarly we can show that

\[
E \left[ \frac{1}{N} \sum_{i=1}^{N} \frac{(1 - Z_i) S_i Y_i}{\hat{f}_i (1 - e_i)} \right] = E[Y^z=0].
\]

Putting these results together we have that

\[
E \left[ \frac{1}{N} \sum_{i=1}^{N} \frac{Z_i S_i Y_i}{\hat{f}_i e_i} - \frac{1}{N} \sum_{i=1}^{N} \frac{(1 - Z_i) S_i Y_i}{\hat{f}_i (1 - e_i)} \right] = E[Y^z=1 - Y^z=0] = \Delta_{PATE}.
\]

In practice, \( e_i \) is unobserved and must be estimated from the data, which leads to the estimator

\[
\hat{\Delta}_{ipw1.svy} = \frac{1}{N} \sum_{i \in S} \frac{Z_i Y_i}{\hat{f}_i \hat{e}_i} - \frac{1}{N} \sum_{i \in S} \frac{(1 - Z_i) Y_i}{\hat{f}_i (1 - \hat{e}_i)}.
\]
Similarly to the proof above, we can show that

\[
E \left[ \sum_{i=1}^{N} \frac{Z_iS_i}{f_i e_i} \right] = E \left[ \sum_{i=1}^{N} \frac{(1 - Z_i)S_i}{f_i(1 - e_i)} \right] = N.
\]

Then a natural estimator that follows is

\[
\hat{\Delta}_{ipw2.svy} = \left( \sum_{i \in S} \frac{Z_i}{f_i \hat{e}_i} \right)^{-1} \sum_{i \in S} \frac{Z_i Y_i}{f_i \hat{e}_i} - \left( \sum_{i \in S} \frac{(1 - Z_i)}{f_i(1 - \hat{e}_i)} \right)^{-1} \sum_{i \in S} \frac{(1 - Z_i)Y_i}{f_i(1 - \hat{e}_i)},
\]

which standardizes the weights compared to (2.7).

The estimators 2.7 and 2.8 can be viewed as belonging to a more general class of estimators similar to their non-survey counterparts. The class can be defined as solutions to the following equations

\[
\sum_{i \in S} \frac{1}{f_i} \left\{ \frac{Z_i(Y_i - \mu_1)}{e_i} + \eta_1 \left( \frac{Z_i - e_i}{e_i} \right) \right\} = 0 , \text{ and}
\]

\[
\sum_{i \in S} \frac{1}{f_i} \left\{ \frac{(1 - Z_i)(Y_i - \mu_0)}{1 - e_i} - \eta_0 \left( \frac{Z_i - e_i}{1 - e_i} \right) \right\} = 0.
\]

Solving for \( \mu_1 \) and \( \mu_0 \) in (2.9) leads to the following equations for any \( \eta_1, \eta_0 \)

\[
\hat{\mu}_1 = \left( \frac{1}{N} \sum_{i \in S} \frac{Z_i}{f_i \hat{e}_i} \right)^{-1} \left\{ \frac{1}{N} \sum_{i \in S} \frac{Z_i Y_i}{f_i \hat{e}_i} - \eta_1 \frac{1}{N} \sum_{i \in S} \frac{1}{f_i} \left( \frac{Z_i}{\hat{e}_i} - 1 \right) \right\}, \text{ and}
\]

\[
\hat{\mu}_0 = \left( \frac{1}{N} \sum_{i \in S} \frac{1 - Z_i}{f_i(1 - e_i)} \right)^{-1} \left\{ \frac{1}{N} \sum_{i \in S} \frac{(1 - Z_i)Y_i}{f_i(1 - e_i)} - \eta_0 \frac{1}{N} \sum_{i \in S} \frac{1}{f_i} \left( 1 - \frac{(1 - Z_i)}{(1 - e_i)} \right) \right\}.
\]

The estimators (2.10) and (2.11) can be separated into individual Horvitz-Thompson estimators (consider each summation). Assuming that each of these individual estimators are design consistent, see 1.3.1 of Fuller [10]; then the class of estimators

\( \hat{\Delta} = \hat{\mu}_1 - \hat{\mu}_0 \) will be design consistent for \( \Delta_{PATE} \).
Again, like the non-survey class of estimators we can attempt to find the values \((\eta_1, \eta_0)\) that minimize the large sample variance; however, in this case we are dealing with a variance that in part comes from the survey design, which complicates the matter. Here we utilize the results from Fuller [10], to derive the large sample variance. This approach uses the asymptotic theory of survey statistics and the M-estimation approach. To extend M-estimation to survey data, we use the \(N\) asymptotic results from [10], where the sample size, \(n\), is a function of the population size, \(N\). We assume the true propensity scores are known.

Consider a sequence of finite populations and their associated probability samples. Keeping with the notation of Fuller [10], assume that the \(N\)th finite population contains \(N\) elements. Let \(y_{iN}\) denote the column vector of characteristics associated with the \(i\)th element of the \(N\)th population. The set of indices for the \(N\)th finite population is defined by

\[
U_N = \{1, 2, \ldots, N\}.
\]  

(2.12)

Then let

\[
F_N = (y_{1N}, y_{2N}, \ldots, y_{NN})
\]  

(2.13)

be the set of vectors containing unit characteristics for the \(N\)th finite population. Let \(S_N\) define the sample selected from the \(N\)th finite population. A sample is defined by a subset of population indices. The number of distinct indices in the sample, the sample size, is given by \(n_N\).

Let \(F_N\) be the \(N\)th finite population in the sequence \(\{F_N\}\), and assume \(\{y_1\}\) is a sequence of \(iid\) random variables with finite fourth moments.
Let \( Y_i = (Z_i, Y_i, e_i) \) and define \( g(Y_i, \theta, \eta_1, \eta_0)' = (\psi_{3.1}(Y_i, \theta, \eta_1), \psi_{3.0}(Y_i, \theta, \eta_0)) \),

where

\[
\psi_{3.1}(Y_i, \theta, \eta_1) = \left\{ \frac{Z_i(Y_i - \mu_1)}{e_i} + \eta_1 \left( \frac{Z_i - e_i}{e_i} \right) \right\}, \quad \text{and} \quad (2.14)
\]

\[
\psi_{3.0}(Y_i, \theta, \eta_0) = \left\{ \frac{(1-Z_i)(Y_i - \mu_0)}{1-e_i} - \eta_0 \left( \frac{Z_i - e_i}{1-e_i} \right) \right\}. \quad (2.15)
\]

Then

\[
\sum_{i \in S} \frac{1}{f_i} g(Y_i, \theta, \eta_1, \eta_0)' = 0 \quad (2.16)
\]
defines our estimating equations. Define \( \theta_N \) as the true population parameter vector for the \( N \)th finite population such that

\[
\sum_{i \in U_N} g(Y_i, \theta_N, \eta_1, \eta_0) = 0. \quad (2.17)
\]

Define \( \theta^o \), the superpopulation parameter, such that \( \theta^o \) satisfies

\[
E \left[ \sum_{i \in U_N} g(Y_i, \theta^o, \eta_1, \eta_0) \right] = 0. \quad (2.18)
\]

Following the results of Theorem 1.3.9 of [10] \([V_\infty(\hat{\theta} - \theta^o)]^{-1/2}(\hat{\theta} - \theta^o)\) converges in distribution to a \( \mathcal{N}(0, I) \) random variable given the existence of certain moments for the superpopulation, a central limit theorem for the Horvitz-Thompson estimator of the total, and that \( g(Y_i, \theta, \eta_1, \eta_0) \) has a continuous second derivative with respect to \( \theta \). We also assume that the estimator \( \hat{\theta} \) is consistent for both \( \theta_N \) and \( \theta^o \). The large sample variance is given by

\[
V \left\{ \hat{\theta} - \theta^o \right\} = A^{-1}(\theta^o)(N^{-1}\Sigma_{gg} + V_{gg,N})A^{-1}(\theta^o), \quad \text{where}
\]

\[
A(\theta^o) = E \left[ -\hat{g}(Y_i, \theta^o, \eta_1, \eta_0) \right],
\]

49
\[ \Sigma_{gg} = E \left[ g(Y_i, \theta^o, \eta_1, \eta_0) g'(x_i, \theta^o, \eta_1, \eta_0) \right], \]

\[ V_{t,gg,N} = V \left\{ N^{-1} \sum_{i \in S} \frac{1}{f_i} g(Y_i, \theta_N, \eta_1, \eta_0) | F_N \right\}, \]

\[ g(Y_i, \theta, \eta_1, \eta_0) = \begin{bmatrix} \frac{Z_i(Y_i - \mu_1)}{e_i} + \eta_1 \left( \frac{Z_i - e_i}{e_i} \right) \\ \frac{(1 - Z_i)(Y_i - \mu_0)}{1 - e_i} - \eta_0 \left( \frac{Z_i - e_i}{1 - e_i} \right) \end{bmatrix}, \text{ and} \]

\[ \dot{g}(Y_i, \theta, \eta_1, \eta_0) = \frac{\partial g(Y_i, \theta, \eta_1, \eta_0)}{\partial \theta} = \begin{bmatrix} -\frac{Z_i}{e_i} & 0 \\ 0 & \frac{-1 - Z_i}{1 - e_i} \end{bmatrix}. \]

Observe that the large sample variance is the sum of two parts, \( \Sigma_{gg} \) and \( V_{t,gg,N} \).

We find that

\[ A(\theta^o) = E \left[ -\dot{g}(Y_i, \theta^o, \eta_1, \eta_0) \right] \]

\[ = -E \begin{bmatrix} -\frac{Z_i}{e_i} & 0 \\ 0 & \frac{-1 - Z_i}{1 - e_i} \end{bmatrix} \]

\[ = I \]

Consider the sampling variance \( V_{t,gg,N} \). For the Horvitz-Thompson estimator of the total \( \hat{T}_y = \sum_{i=1}^n \frac{y_i}{f_i} \), its variance can be written as

\[ V(\hat{T}_y) = \sum_{i=1}^N \left( \frac{y_i}{f_i} \right)^2 (1 - f_i)(f_i) + \sum_{i=1}^N \sum_{j \neq i}^N (f_{ij} - f_i f_j) \frac{y_i y_j}{f_i f_j}, \quad (2.19) \]

where \( f_{ij} \) is the joint probability of inclusion of observations \( i \) and \( j \) in the sample.

This can further be written as
\[
V(\hat{T}_y) = \sum_{i=1}^{N} \left( \frac{y_i}{f_i} \right)^2 (1 - f_i)(f_i) + \sum_{i=1}^{N} \sum_{j \neq i} (f_{ij} - f_i f_j) \frac{y_i y_j}{f_i f_j}
\]

\[
= \sum_{i=1}^{N} y_i^2 \frac{1 - f_i}{f_i} + \sum_{i=1}^{N} \sum_{j \neq i} (f_{ij} - f_i f_j) \frac{y_i y_j}{f_i f_j}
\]

\[
= \sum_{i=1}^{N} y_i^2 \frac{1 - f_i}{f_i} - \sum_{i=1}^{N} y_i^2 + \sum_{i=1}^{N} \sum_{j \neq i} (f_{ij} - f_i f_j) \frac{y_i y_j}{f_i f_j}
\]

\[
= (N)E \left[ \frac{Y^2}{f} \right] - (N)E[Y^2] + \sum_{i=1}^{N} \sum_{j \neq i} (f_{ij} - f_i f_j) \frac{y_i y_j}{f_i f_j}
\]

Consider an approximation of the variance that only consists of the first two terms above.

\[
V(\hat{T}_y) \approx (N)E \left[ \frac{Y^2}{f} \right] - (N)E[Y^2].
\]  

(2.20)

When the sampling design is such that \( f_{ij} = f_i f_j \) for all \( i \) and \( j \), then this last term will be zero anyways. This could happen with a simple stratification design. Units within a stratum will be independently sampled and units between strata will also be independently sampled, which implies the joint probability of selection of two units is equal to the product of the two individual probabilities. When the sampling design is such that \( f_{ij} \neq f_i f_j \), such as in a cluster design, the approximation will not be as good. The primary reason for this approximation is practicality and ease of use for this estimator.

Recall that we are trying to minimize the variance of \( \hat{\Delta} = \hat{\mu}_1 - \hat{\mu}_0 \). Following the work of Lunceford and Davidian in their derivation of \( \hat{\Delta}_{ipw3} \) [27], we separately minimize the variances of \( \hat{\mu}_1 \) and \( \hat{\mu}_0 \). This minimizes the variance of \( \hat{\Delta} \) assuming \( \hat{\mu}_1 \) and \( \hat{\mu}_0 \) are independent. Applying the variance approximation above to \( V_{t,gg,N} \) yields
\[ V_{t,gg,N}(\psi_{3.1}) \approx N^{-1}E \left[ \frac{\psi_{3.1}(Y_i, \theta, \eta_1)}{f_i} \right] - N^{-1}E[\psi_{3.1}(Y_i, \theta, \eta_1)^2], \text{ and} \]

\[ V_{t,gg,N}(\psi_{3.0}) \approx N^{-1}E \left[ \frac{\psi_{3.0}(Y_i, \theta, \eta_0)}{f_i} \right] - N^{-1}E[\psi_{3.0}(Y_i, \theta, \eta_0)^2]. \]

Recall that \( \Sigma_{gg} = E \left[ g(Y_i, \theta^o, \eta_1)g'(x_i, \theta^o, \eta_1, \eta_0) \right] \) and that

\[ V \left\{ \hat{\theta} - \theta^o \right\} = A^{-1}(\theta^o)(N^{-1}\Sigma_{gg} + V_{t,gg,N})A'^{-1}(\theta^o). \]

Observe that \( N^{-1}E[\psi_{3.1}(Y_i, \theta, \eta_1)^2] \) and \( N^{-1}E[\psi_{3.0}(Y_i, \theta, \eta_0)^2] \) will both cancel so that

\[ V \left\{ \hat{\theta} - \theta^o \right\} \approx \left[ N^{-1}E \left[ \frac{\psi_{3.1}(Y_i, \theta, \eta_1)}{f_i} \right] \right] \frac{a}{N^{-1}E \left[ \frac{\psi_{3.0}(Y_i, \theta, \eta_0)}{f_i} \right]}, \]

where \( a \) is the covariance term that will not play a role in the minimization. Next we evaluate \( E \left[ \frac{\psi_{3.1}(Y_i, \theta, \eta_1)}{f_i} \right] \). We see that

\[ E \left[ \frac{\psi_{3.1}(Y_i, \theta, \eta_1)}{f_i} \right] = E \left[ \frac{1}{f_i} \left\{ \left( \frac{Z_i(Y_i - \mu_1)}{e_i} \right)^2 + \eta_1^2 \left( \frac{Z_i - e_i}{e_i} \right)^2 + 2\eta_1 \frac{Z_i(Y_i - \mu_1)}{e_i} \left( \frac{Z_i - e_i}{e_i} \right) \right\} \right]. \]

In order to find the value of \( \eta_1 \) that gives us the minimum variance, we take the derivative with respect to \( \eta_1 \) and set this equal to zero.

\[ \frac{\partial E \left[ \frac{\psi_{3.1}(Y_i, \theta, \eta_1)}{f_i} \right]}{\partial \eta_1} = E \left[ \frac{1}{f_i} \left\{ 2\eta_1 \left( \frac{Z_i - e_i}{e_i} \right)^2 + 2Z_i(Y_i - \mu_1) \left( \frac{Z_i - e_i}{e_i} \right) \right\} \right]. \]

After some algebra, solving for \( \eta_1 \) gives the solution
\[
\eta_1 = -E \left[ \frac{1}{f_i} \left( \frac{Z_i(Y_i - \mu_1)}{e_i^2} \right) \right] / E \left[ \frac{1}{f_i} \left( \frac{Z_i - e_i}{e_i} \right)^2 \right]. \tag{2.21}
\]

Similarly we evaluate \( E \left[ \frac{\psi_\eta(Y, \theta, \eta_0)^2}{f_i} \right] \). We find that

\[
E \left[ \frac{\psi_\eta(Y, \theta, \eta_0)^2}{f_i} \right] = E \left[ \frac{1}{f_i} \left\{ \left( \frac{(1 - Z_i)(Y_i - \mu_0)}{(1 - e_i)} \right)^2 + \eta_0^2 \left( \frac{Z_i - e_i}{1 - e_i} \right)^2 
+ 2\eta_0 \frac{(1 - Z_i)(Y_i - \mu_0)}{(1 - e_i)^2} \left( \frac{Z_i - e_i}{e_i} \right) \right\} \right].
\]

Next, we take the derivative of this term with respect to \( \eta_0 \), set it equal to zero, and solve for \( \eta_0 \) in order to find the value that minimizes the variance.

\[
\frac{\partial E \left[ \frac{\psi_\eta(Y, \theta, \eta_0)^2}{f_i} \right]}{\partial \eta_0} = E \left[ \frac{1}{f_i} \left\{ 2\eta_0 \left( \frac{Z_i - e_i}{1 - e_i} \right)^2 + 2 \frac{(1 - Z_i)(Y_i - \mu_0)}{1 - e_i} \left( \frac{Z_i - e_i}{e_i} \right) \right\} \right].
\]

Setting this equal to zero and solving for \( \eta_0 \) gives the solution

\[
\eta_0 = -E \left[ \frac{1}{f_i} \left( \frac{(1 - Z_i)(Y_i - \mu_0)}{(1 - e_i)^2} \right) \right] / E \left[ \frac{1}{f_i} \left( \frac{Z_i - e_i}{1 - e_i} \right)^2 \right]. \tag{2.22}
\]

These constants need to be estimated, which we can do by solving additional estimating equations. These equations are given by

\[
\sum_{i \in S} \frac{1}{f_i} \psi_{\eta_1}(Y_i, \theta, \eta_1) = 0, \quad \text{and}
\]

\[
\sum_{i \in S} \frac{1}{f_i} \psi_{\eta_0}(Y_i, \theta, \eta_0) = 0, \tag{2.23}
\]

where

\[
\psi_{\eta_1}(Y_i, \theta, \eta_1) = \eta_1 \left( \frac{Z_i - e_i}{e_i^2} \right)^2 + \frac{Z_i(Y_i - \mu_1)}{e_i^2}, \quad \text{and} \tag{2.24}
\]

53
\[
\psi_{\eta_0}(Y_i, \theta, \eta_0) = \eta_0 \left( \frac{Z_i - \epsilon_i}{1 - \epsilon_i} \right)^2 + \frac{(1 - Z_i)(Y_i - \mu_0)}{(1 - \epsilon_i)^2}
\] (2.25)

In practice we also estimate the propensity scores. Assume that we estimate the propensity scores using the logistic regression model \(e(X, \beta) = \{1 + \exp(-X'\beta)\}^{-1}\) and estimate \(\beta\) by weighted maximum likelihood by solving

\[
\sum_{i \in S} \psi_{\beta}(Z_i, X_i, \beta) = \sum_{i \in S} \frac{1}{f_i} \frac{Z_i - e(X_i, \beta)}{e(X_i, \beta)(1 - e(X_i, \beta))} \frac{\partial}{\partial \beta} e(X_i, \beta) = 0.
\] (2.26)

Solving (2.16), (2.23), and (2.26) jointly yields

\[
\hat{\Delta}_{ipw3.svy} = \left\{ \sum_{i \in S} \frac{1}{f_i} \left( \frac{Z_i}{\hat{\epsilon}_i} \right) \left( 1 - \frac{C_1}{f_i \hat{\epsilon}_i} \right) \right\}^{-1} \sum_{i \in S} \frac{1}{f_i} \left( \frac{Z_i Y_i}{\hat{\epsilon}_i} \right) \left( 1 - \frac{C_1}{f_i \hat{\epsilon}_i} \right) - \left\{ \sum_{i \in S} \frac{1}{f_i} \left( 1 - \frac{Z_i}{\hat{\epsilon}_i} \right) \left( 1 - \frac{C_0}{f_i (1 - \hat{\epsilon}_i)} \right) \right\}^{-1} \times \sum_{i \in S} \frac{1}{f_i} \left( \frac{(1 - Z_i) Y_i}{1 - \hat{\epsilon}_i} \right) \left( 1 - \frac{C_0}{f_i (1 - \hat{\epsilon}_i)} \right)
= \hat{\mu}_{1,ipw3.svy} - \hat{\mu}_{0,ipw3.svy},
\] where

\[
C_1 = \sum_{i \in S} \frac{1}{f_i} \left( \frac{Z_i - \hat{\epsilon}_i}{\hat{\epsilon}_i} \right) / \sum_{i \in S} \frac{1}{f_i} \left( \frac{Z_i - \hat{\epsilon}_i}{\hat{\epsilon}_i} \right)^2, \text{ and}
\]

\[
C_0 = -\sum_{i \in S} \frac{1}{f_i} \left( \frac{Z_i - \hat{\epsilon}_i}{1 - \hat{\epsilon}_i} \right) / \sum_{i \in S} \frac{1}{f_i^2} \left( \frac{Z_i - \hat{\epsilon}_i}{1 - \hat{\epsilon}_i} \right)^2.
\]

### 2.4 Estimating the Propensity Score

In practice we never know or observe the true propensity score and must estimate it from the observed data. Typically we assume the propensity score follows a parametric model such as a logistic regression model. When dealing with complex survey data we have the added option of utilizing the survey weights in the model. If
\( e(X_i, \beta) = (1 + exp(-X_i'\beta))^{-1} \) is our logistic regression model, then the coefficients \( \beta \) can be estimated by maximizing the log-likelihood function

\[
\sum_{i \in S} \left[Z_i \log(e(X_i, \beta)) + (1 - Z_i) \log(e(X_i, \beta))\right].
\] (2.28)

When including the survey weights in the estimation process, we estimate \( \beta \) by maximizing the pseudo log-likelihood function

\[
\sum_{i \in S} \frac{1}{f_i} \left[Z_i \log(e(X_i, \beta)) + (1 - Z_i) \log(e(X_i, \beta))\right].
\] (2.29)

The common practice in the literature is that the survey weights are not needed to weight the log-likelihood function as in (2.29). Some argue that if the sampling weights or the survey design variables include information on the treatment selection process, then they should be included in the model as covariates, but otherwise are unneeded [9]. We argue, however, that the survey-weighted logistic regression (2.29) is the correct estimating procedure for the propensity scores in the complex survey setting.

The ultimate goal of estimating and using propensity scores is to balance the distributions of selected covariates between treatment and control groups, or at least to reduce the imbalance. In this regard, we do not care much about model diagnostics and can include every feasible interaction in the model if it helps. When estimating the population average treatment effect using complex survey data, we need to make sure we preserve population representativeness. Also, the estimating equation technique that we are using for the weighting estimators to estimate the variance rely on the fact that the sampling weights are part of the estimating equation for the logistic regression estimating equations. Any variance estimates using propensity scores estimated without the sampling weights would have to be approached differently than
we will present, else the variance estimates are not technically sound. Theoretical arguments aside, if estimating the propensity scores using a survey-weighted logistic regression improves the balance after weighting or stratifying over an unweighted logistic regression, there is no reason not to use it.

There are two reasons why estimating the propensity score via a survey-weighted logistic regression will lead to better balance. First, in our theoretical work (Section 2.1) we are assuming that \( e_i = E[Z|X_i] \). We take advantage of this conditional expectation to cancel out terms, and in order to prove unbiasedness the propensity score must reflect the population distribution, not the sample distribution. Therefore, we should estimate our propensity scores in a way that reflects the population distribution. Secondly, both stratified and weighting estimators utilize the sampling weights as part of a weighted average. We can think of this in sampling terms as counting for other un-sampled units in the population. Similarly the sampled unit should count for the same number of sampled units when estimating the propensity score.

We expect stratified estimators not to be greatly affected by estimating the propensity score via a survey-weighted logistic regression versus an unweighted logistic regression because a small difference in the distribution of propensity scores should not greatly affect which units are grouped together. However, weighting estimators are likely to be sensitive to even small changes in the estimated propensity scores. Even when there is only a small difference between the estimated propensity scores using survey-weighted logistic regression versus an unweighted logistic regression we expect substantial differences in the balance and ultimately the estimate from weighting estimators.

Ultimately, applying the weights to the logistic regression is very simple and should be used if it improves balance. If design variables are found to be predictive of
treatment assignment, they need to be included in the logistic regression model, but otherwise need not be included. We previously characterized design variables that are predictive of treatment assignment as $X^{ZS}$. This is a separate issue from the assumption that $Z \perp S|X$.

2.5 Evaluating Balance

One of the goals of using propensity scores is to balance the observed pre-treatment confounders between the treated and control groups. Checking the balance should be an intermediate step before we estimate the treatment effect using any of the estimators. We should think about the balance before propensity score adjustment and afterwards. In a non-complex survey context, the before-balance is the comparison of the sample distributions of treated and control units. This is often evaluated by the standardized difference \cite{4} \cite{5}. The percent standardized difference is written as:

$$\text{std.diff}(x)_{\text{cont}} = 100 \times \frac{\text{abs}(\bar{x}_{z=1} - \bar{x}_{z=0})}{\sqrt{(s^2_{z=1} + s^2_{z=0})/2}}, \quad (2.30)$$

for a continuous variable $x$, with sample treatment and control sample means $\bar{x}_{z=1}$ and $\bar{x}_{z=0}$ and sample treatment and control standard deviations $s_{z=1}$ and $s_{z=0}$. For a categorical variable with $k$ categories, treat the categories as indicator variables and calculate $k - 1$ differences using:

$$\text{std.diff}(x)_{\text{categ}} = 100 \times \frac{\text{abs}(\hat{p}_{x,z=1} - \hat{p}_{x,z=0})}{\sqrt{((\hat{p}_{x,z=1})(1 - \hat{p}_{x,z=1}) + (\hat{p}_{x,z=0})(1 - \hat{p}_{x,z=0})/2}}, \quad (2.31)$$

where $\hat{p}_{x,z=1}$ and $\hat{p}_{x,z=0}$ are sample proportions.

In the complex survey context, where we are estimating the population effect, the before-balance is the comparison of the estimated population distributions of treated
and control units. The estimated population distributions are the weighted distributions of the sample units, using sampling weights. This is the correct comparison because if we want to make an estimate of the population average treatment effect, we would compare the weighted treatment and control averages. Simply replace $\bar{x}_{z=1}$, $\bar{x}_{z=0}$, $s_{z=1}$, $s_{z=0}$, $\hat{p}_{x,z=1}$, and $\hat{p}_{x,z=0}$ with their weighted versions in (2.30) and (2.31). We use

$$s^2_{\text{weight}} = \frac{1}{W_1 - 1} \sum_{i=1}^{n} w_i (x_i - \bar{x}_w)^2$$

(2.32)

$$W_1 = \sum_{i=1}^{n} w_i \quad \bar{x}_w = \frac{1}{W_1} \sum_{i=1}^{n} w_i x_i$$

to estimate the weighted variance using weights $w_i, i = 1, \ldots, n$. In this case $w_i = 1/f_i$. This estimates the variance from the empirical probability mass function [26].

The balance after propensity score adjustment is estimator specific. Starting in the non-complex survey context, if we are estimating the treatment effect by paired matches, then we compare the balance between the matches. If we are using a stratified estimator, then we compare the balance inside each propensity score stratum. If we are using a weighting estimator, we compare the balance of the pseudo-population, that is the weighted distribution of units [16]. Now in the complex survey context, the same logic should hold true. For the estimators $\hat{\Delta}_{ipw1,svy}$ and $\hat{\Delta}_{ipw2,svy}$ replace $\bar{x}_{z=1}$, $\bar{x}_{z=0}$, $s_{z=1}$, $s_{z=0}$, $\hat{p}_{x,z=1}$, and $\hat{p}_{x,z=0}$ with their weighted versions in (2.30) and (2.31) using weights $w_i = z_i/(f_i \hat{e}_i) + (1 - z_i)/(f_i (1 - \hat{e}_i))$. For $\hat{\Delta}_{ipw3,svy}$ use weights equal to

$$w_i = \frac{z_i}{f_i \hat{e}_i} \left( 1 - \frac{C_1(i)}{f_i \hat{e}_i} \right) + \frac{(1 - z_i)}{f_i (1 - \hat{e}_i)} \left( 1 - \frac{C_0(i)}{f_i (1 - \hat{e}_i)} \right).$$
For the survey-adjusted stratified estimator $\hat{\Delta}_{s.svy}$, we evaluate the after-balance in each propensity score stratum. Simply apply (2.30) and (2.31) with weights $w_i = 1/f_i$ locally to each propensity score stratum.

### 2.6 Variance Estimators

We propose variance estimators for both the stratified estimator (2.5) and weighting estimators (2.7, 2.8, 2.27) that utilize survey design features.

#### 2.6.1 Variance of Stratification Estimators

As with the variance estimators (1.33) and (1.35), we continue to assume independence in the estimators across propensity score strata in the survey data context. Even when this is assumption is not true, when the dependence between units across propensity score strata is small, this method should still provide a close approximation of the variance. There are multiple possible approximations for the variance of the propensity score stratum estimates based on further assumptions about the independence of estimators within propensity score strata. First let us begin with the most complex version of the variance estimate by considering the covariance between all terms. We use a Taylor series linearization to approximate the variance. The estimated variance of $\hat{\Delta}_{s.svy}$ is given by

$$
\hat{\text{var}}(\hat{\Delta}_{s.svy}) = \sum_{k=1}^{K} \hat{\text{var}}(\hat{\theta}_k),
$$

where

$$
\hat{\theta}_k = \frac{\hat{N}_k}{N} \left[ \hat{N}_k^{-1} \sum_{i \in S} \frac{Z_i I(\hat{e}_i \in Q_k)Y_i}{f_i} - \hat{N}_{k0}^{-1} \sum_{i \in S} \frac{(1 - Z_i)I(\hat{e}_i \in Q_k)Y_i}{f_i} \right].
$$

The estimator $\hat{\theta}_k$ can be written as a function of estimates of population totals...
\[ \hat{\theta}_k = \frac{1}{N} \left( \hat{U}_k + \hat{V}_k \right) \left( \frac{\hat{A}_k}{\hat{U}_k} - \frac{\hat{B}_k}{\hat{V}_k} \right), \]

where \( \hat{U}_k = \hat{N}_{k1}, \hat{V}_k = \hat{N}_{k0}, \hat{A}_k = \sum_{i \in \mathcal{S}} \frac{Z_i(\hat{e}_i \in \mathcal{Q}_k)Y_i}{f_i}, \) and \( \hat{B}_k = \sum_{i \in \mathcal{S}} \frac{(1-Z_i)(\hat{e}_i \in \mathcal{Q}_k)Y_i}{f_i}. \)

For the moment, let us drop the subscripts and let

\[ f(W) = (U + V) \left( \frac{A}{U} - \frac{B}{V} \right). \]

The partial derivatives of \( f(\cdot) \) are

\[ \frac{\partial f(W)}{\partial U} = \left( \frac{A}{U} - \frac{B}{V} \right) + (U + V) \left( \frac{-A}{U^2} \right), \]

\[ \frac{\partial f(W)}{\partial V} = \left( \frac{A}{U} - \frac{B}{V} \right) + (U + V) \left( \frac{B}{V^2} \right), \]

\[ \frac{\partial f(W)}{\partial A} = (U + V) \frac{1}{U} = \left( 1 + \frac{V}{U} \right), \text{ and} \]

\[ \frac{\partial f(W)}{\partial B} = (U + V) \frac{-1}{V} = - \left( 1 + \frac{U}{V} \right). \]

The general formula for variance approximation is

\[ V(\hat{\theta}) = \sum_i \sum_k \frac{\partial f(W)}{\partial \hat{W}_i} \frac{\partial f(W)}{\partial \hat{W}_k} V(\hat{W}_i, \hat{W}_k). \]  \hspace{1cm} (2.34)

This leads to
\[ V_1(\hat{\theta}) = \left( \frac{\partial f(W)}{\partial U} \right)^2 Var(U) + \left( \frac{\partial f(W)}{\partial V} \right)^2 Var(V) \]
\[ + \left( \frac{\partial f(W)}{\partial A} \right)^2 Var(A) + \left( \frac{\partial f(W)}{\partial B} \right)^2 Var(B) \]
\[ + 2 \left( \frac{\partial f(W)}{\partial U} \frac{\partial f(W)}{\partial V} \right) Cov(U,V) + 2 \left( \frac{\partial f(W)}{\partial U} \frac{\partial f(W)}{\partial A} \right) Cov(U,A) \]
\[ + 2 \left( \frac{\partial f(W)}{\partial B} \frac{\partial f(W)}{\partial V} \right) Cov(U,B) + 2 \left( \frac{\partial f(W)}{\partial V} \frac{\partial f(W)}{\partial A} \right) Cov(V,A) \]
\[ + 2 \left( \frac{\partial f(W)}{\partial V} \frac{\partial f(W)}{\partial B} \right) Cov(V,B) + 2 \left( \frac{\partial f(W)}{\partial A} \frac{\partial f(W)}{\partial B} \right) Cov(A,B). \] (2.35)

For the partial derivatives, we use the formulas above and plug in the estimates of the totals from the data. The variance and covariance estimate formulas are survey design dependent. Wolter Chapter 1 [46] gives examples of many of these formulas.

For example, assume we have a stratified sampling design. Changing the notation for the moment, let \( y_{ji} \) be the value for the \( i \)th observation from the \( j \)th sampling stratum, and \( w_{ji} \) be the sampling weight for that same unit. Assume there are \( J \) strata, with a sample of \( n_j \) taken from stratum \( j \). Say we are trying to estimate the total for some variable \( Y \). Denote the estimate of this total by \( \hat{Y}_T \). An estimate of the variance of \( \hat{Y}_T \) is

\[
\hat{V}(\hat{Y}_T) = \sum_{j=1}^{J} \left( \frac{n_j}{n_j - 1} \right) \sum_{i=1}^{n_j} w_{ji}^2 (y_{ji} - \bar{y}_j)^2,
\]

\[
\bar{y}_j = \frac{\sum_{i=1}^{n_j} w_{ji} y_{ji}}{\sum_{i=1}^{n_j} w_{ji}}.
\]

Let \( x_{ji} \) and \( \bar{x}_j \) be similarly defined as above for some variable \( X \). Then the covariance of the estimated totals of \( X \) and \( Y \) is estimated by

\[
\hat{COV}(\hat{Y}_T, \hat{X}_T) = \sum_{k=1}^{K} \left( \frac{n_j}{n_j - 1} \right) \sum_{i=1}^{n_j} w_{ji}^2 (y_{ji} - \bar{y}_j) (x_{ji} - \bar{x}_j).
\]

61
For many approximations of the stratified estimator the treatment and control estimators are considered independent, such as in (1.33), which implies that their covariance is zero. This reduces (2.35) to

\[
V(\hat{\theta}) = \left(\frac{\partial f(W)}{\partial U}\right)^2 Var(U) + \left(\frac{\partial f(W)}{\partial V}\right)^2 Var(V) \\
+ \left(\frac{\partial f(W)}{\partial A}\right)^2 Var(A) + \left(\frac{\partial f(W)}{\partial B}\right)^2 Var(B) \\
+ 2 \left(\frac{\partial f(W)}{\partial U} \frac{\partial f(W)}{\partial A}\right) Cov(U, A) + 2 \left(\frac{\partial f(W)}{\partial V} \frac{\partial f(W)}{\partial B}\right) Cov(V, B).
\]

Therefore, depending on our approximation \(V(\hat{\theta}) = (2.35)\) or \(2.36\), we estimate the variance of \(\hat{\Delta}_{s,svy}\) by

\[
v\hat{\text{ar}}(\hat{\Delta}_{s,svy}) = \sum_{k=1}^{K} \frac{1}{N^2} \hat{V}(\hat{\theta}_k),
\]

where \(V(\hat{\theta})\) is defined as above, but with estimates of the respective totals of \(U_k, V_k, A_k,\) and \(B_k\) and variance and covariance estimates based on survey-design-specific formulas.

\subsection{2.6.2 Variance of Weighting Estimators}

We derive the variance of the proposed weighting estimators using the asymptotic theory of survey statistics and the M-estimation approach. Variance approximation of an M-estimator for non-survey data is discussed by Lunceford and Davidian [27].

To extend the M-estimation to survey data, we use the \(N\) asymptotic results from Fuller [10], where the sample size, \(n\), is a function of the population size, \(N\).

Let \(e(X_i, \beta) = (1 + exp(-X_i'\beta))^{-1}\) specify our logistic regression model and \(e_\beta = \frac{\partial}{\partial \beta}(X_i, \beta)\). Then define \(g(Y_i, \theta)' = (\psi_1(Y_i, \theta), \psi_\beta(Y_i, \theta))\), where
\[
\psi_1(Y_i, \theta) = \left\{ \left( \frac{Z_i Y_i}{e(X_i, \beta)} - \frac{(1 - Z_i)Y_i}{(1 - e(X_i, \beta))} \right) - \Delta \right\}, \text{ and}
\]
\[
\psi_{\beta}(Y_i, \theta) = \left( \frac{Z_i - e(X_i, \beta)}{e(X_i, \beta)(1 - e(X_i, \beta))} \right) e_{\beta}.
\]

The function \( g_1(Y_i, \theta) \) is a vector function of length \( j + 1 \), where \( j \) is the length of the parameter vector \( \beta \).

Setting \( \sum_{i \in S} \frac{1}{n_i} g_1(Y_i, \theta)' = 0 \) and solving for \( \theta \) implies 1) \( \hat{\Delta} = \hat{\Delta}_{ipw1.svy} \), and 2) \( \hat{\beta} \) is the solution to maximizing the pseudo log-likelihood function, as in equation (2.29). If we instead define

\[
\psi_2(Y_i, \theta) = \left( \frac{Z_i - \frac{Y_i - \mu_1}{e(X_i, \beta)}}{1 - Z_i} \right),
\]

then setting \( g_2(Y_i, \theta)' = (\psi_2'(Y_i, \theta), \psi_{\beta}'(Y_i, \theta)) \) and solving \( \sum_{i \in S} \frac{1}{n_i} g_2(Y_i, \theta)' = 0 \) for \( \theta \) implies \( \hat{\Delta} = \hat{\mu}_{1,ipw2.svy} - \hat{\mu}_{0,ipw2.svy} = \hat{\Delta}_{ipw2.svy} \) as well as the estimates \( \hat{\beta} \).

Finally setting

\[
g_3(Y_i, \theta)' = (\psi_{3,1}(Y_i, \theta), \psi_{3,0}(Y_i, \theta), \psi_{\eta_1}(Y_i, \theta), \psi_{\eta_0}(Y_i, \theta), \psi_{\beta}'(Y_i, \theta))
\]

from equations (2.14), (2.15), (2.24), and (2.25) and solving \( \sum_{i \in S} \frac{1}{n_i} g_3(Y_i, \theta)' = 0 \) for \( \theta \) implies \( \hat{\Delta} = \hat{\mu}_{1,ipw3.svy} - \hat{\mu}_{0,ipw3.svy} = \hat{\Delta}_{ipw3.svy} \) as well as the estimates \( \hat{\beta} \).

Following the results of Theorem 1.3.9 and Corollary 1.3.9.1 of Fuller [10], given several assumptions, \( V_{\infty}(\hat{\theta} - \theta) \sim N(0, I) \) random variable. These assumptions are the existence of certain moments for the superpopulation, a central limit theorem for the Horvitz-Thompson estimator of the total, and that \( g(Y_i, \theta) \) has a continuous second derivative with respect to \( \theta \). We also assume that the estimator \( \hat{\theta} \) is consistent for both \( \theta_N \) and \( \theta^o \). For \( k = 1, 2, 3, \ldots \),
\( \hat{\Delta}_{ipw1.svy}, \hat{\Delta}_{ipw2.svy}, \) and \( \hat{\Delta}_{ipw3.svy} \) respectively, we estimate the variance covariance matrix \( V_{\infty}(\theta_k - \theta^o) \) by

\[
\hat{V}_{\infty}(\hat{\theta}_k - \theta^o) = \hat{A}_k^{-1} \hat{B}_k \hat{A}_k^{-1}, \quad \text{where}
\]

\[
\hat{A}_1^{-1} = \begin{bmatrix}
1 & -\hat{H}_{\beta,1} \hat{E}_{\beta\beta}^{-1} \\
0 & \hat{E}_{\beta\beta}^{-1}
\end{bmatrix}, \quad \hat{A}_2^{-1} = \begin{bmatrix}
1 & 0 & -\hat{H}_{\beta,2,1} \hat{E}_{\beta\beta}^{-1} \\
0 & 1 & \hat{H}_{\beta,2,0} \hat{E}_{\beta\beta}^{-1} \\
0 & 0 & \hat{E}_{\beta\beta}^{-1}
\end{bmatrix}.
\]

\[
\hat{A}_3^{-1} = \begin{bmatrix}
1 & 0 & 0 & 0 & -\hat{H}_{\beta,3,1} \hat{E}_{\beta\beta}^{-1} \\
0 & 1 & 0 & 0 & \hat{H}_{\beta,3,0} \hat{E}_{\beta\beta}^{-1} \\
0 & 0 & 0^* & 0 & 0^* \\
0 & 0 & 0 & 0^* & 0^* \\
0 & 0 & 0 & 0 & \hat{E}_{\beta\beta}^{-1}
\end{bmatrix}.
\]

Note that in the above matrix, \( 0^* \) are not necessarily zero vectors, but they are unimportant to our ultimate variance calculation because we only need a subset of the variance matrix. We thus do not explicitly write them out for the sake of clarity.

\[
\hat{E}_{\beta\beta} = \hat{N}^{-1} \sum_{i \in S} \frac{1}{f_i} \hat{e}_i (1 - \hat{e}_i) X_i X_i', \quad \hat{N} = \sum_{i \in S} \frac{1}{f_i},
\]

\[
\hat{H}_{\beta,1} = \hat{N}^{-1} \sum_{i \in S} \frac{1}{f_i} \left( \frac{Z_i Y_i (1 - \hat{e}_i)}{\hat{e}_i} + \frac{(1 - Z_i) (Y_i) \hat{e}_i}{(1 - \hat{e}_i)} \right) X_i,
\]

\[
\hat{H}_{\beta,2,1} = \hat{N}^{-1} \sum_{i \in S} \frac{1}{f_i} \left( \frac{Z_i (Y_i - \hat{\mu}_1) (1 - \hat{e}_i)}{\hat{e}_i} \right) X_i,
\]

\[
\hat{H}_{\beta,2,0} = \hat{N}^{-1} \sum_{i \in S} \frac{1}{f_i} \left( \frac{(1 - Z_i) (Y_i - \hat{\mu}_0) \hat{e}_i}{(1 - \hat{e}_i)} \right) X_i,
\]

\[
\hat{H}_{\beta,3,1} = \hat{N}^{-1} \sum_{i \in S} \frac{1}{f_i} \left( \frac{Z_i (Y_i - \hat{\mu}_1 + \hat{\eta}_1) (1 - \hat{e}_i)}{\hat{e}_i} \right) X_i,
\]
\[ \hat{\theta}_{3,0} = \hat{N}^{-1} \sum_{i \in S} \frac{1}{f_i} \left( (1 - Z_i)(Y_i - \hat{\mu}_0 + \hat{\eta}_0)e_i \right) X_i, \]

\[ \hat{\eta}_1 = -\sum_{i \in S} \frac{1}{f_i^2} \left( \frac{Z_i(Y_i - \mu_1)}{e_i^2} \right) / \sum_{i \in S} \frac{1}{f_i^2} \left( \frac{Z_i - e_i}{e_i} \right)^2, \]

\[ \hat{\eta}_0 = -\sum_{i \in S} \frac{1}{f_i^2} \left( \frac{(1 - Z_i)(Y_i - \mu_0)}{(1 - e_i)^2} \right) / \sum_{i \in S} \frac{1}{f_i^2} \left( \frac{Z_i - e_i}{1 - e_i} \right)^2, \]

\[ \hat{B}_k = \hat{N}^{-1} \Sigma_{ggk} + \hat{N}^{-2} \Sigma_{tot_k}, \]

\[ \Sigma_{ggk} = \hat{N}^{-1} \sum_{i \in S} \frac{1}{f_i} g_k(Y_i, \hat{\theta})g_k(Y_i, \hat{\theta})', \]

\[ \Sigma_{tot_k} = \hat{V} \left( \sum_{i \in S} \frac{1}{f_i} g_k(Y_i, \hat{\theta}) | \mathcal{F}_N \right) \]

make up the estimated sampling variance matrix of the survey total of \( g_k(Y_i, \hat{\theta}) \). This variance can be estimated by a design-specific estimator using statistical software.

For inference about \( \Delta \), we only care about a portion of \( \hat{V}_\infty(\hat{\theta}_k - \theta^o) \). The estimated variances of \( \hat{\Delta}_{ipw1.svy} \), \( \hat{\Delta}_{ipw2.svy} \), and \( \hat{\Delta}_{ipw3.svy} \) can be computed by

\[ \hat{V}(\hat{\Delta}_{ipw1.svy}) = \begin{bmatrix} 1 & 0' \end{bmatrix} \hat{A}_1^{-1} \hat{B}_1 \hat{A}_1^{-1} \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \]

\[ \hat{V}(\hat{\Delta}_{ipw2.svy}) = \begin{bmatrix} 1 & 0' \end{bmatrix} \hat{A}_2^{-1} \hat{B}_2 \hat{A}_2^{-1} \begin{bmatrix} 1 \\ -1 \\ 0 \end{bmatrix}, \]

\[ \hat{V}(\hat{\Delta}_{ipw3.svy}) = \begin{bmatrix} 1 & 0' \end{bmatrix} \hat{A}_3^{-1} \hat{B}_3 \hat{A}_3^{-1} \begin{bmatrix} 1 \\ -1 \\ 0 \end{bmatrix}. \]
2.7 Estimators for $\Delta_{PATT}$

Just like propensity score estimators in the non-survey setting, we can make adjustments to our propensity score estimators in the complex survey setting so that they estimate $\Delta_{PATT}$ instead of $\Delta_{PATE}$. Additionally, estimators of $\Delta_{PATC}$ are possible. Estimators of $\Delta_{PATC}$ will be very similar to the following equations, or they can be created simply by swapping the definition of treated and control. Instead of thinking about estimating the treatment effect in the superpopulation consisting of pairs of potential outcomes for each unit in the population, we instead want to estimate the treatment effect in a superpopulation consisting of pairs of potential outcomes for each treated (or control) unit in the population. We write out the point estimators and describe the variance estimators. For sake of time and space we do not explicitly derive either. Their derivations and motivations will be similar to what we have already described for the stratification and weighting estimators of $\Delta_{PATE}$.

2.7.1 Stratification Estimators

Only small changes are needed to the stratification estimator. Mainly, we substitute $\hat{N}_1$ for $N$ and $\hat{N}_{k1}$ for $\hat{N}_{k1} + \hat{N}_{k0}$. The within-strata treatment effect estimates are now averaged with respect to the estimated number of treated observations in each stratum. The estimator is given by

$$\hat{\Delta}^t_{k,svy} = \sum_{k=1}^{K} \frac{\hat{N}_{k1}}{\hat{N}_1} \left[ \frac{1}{N_{k1}} \sum_{i=1}^{N} \frac{Z_i S_i I(\hat{e}_i \in Q_k)Y_i}{f_i} - \frac{1}{N_{k0}} \sum_{i=1}^{N} \frac{(1-Z_i) S_i I(\hat{e}_i \in Q_k)Y_i}{f_i} \right],$$

(2.42)

where
\[ \hat{N}_{k1} = \sum_{i \in S} Z_i I(e_i \in Q_k) \frac{1}{f_i} \text{ and } \hat{N}_1 = \sum_{i \in S} Z_i \frac{1}{f_i}. \]

The estimated variance of our \( \hat{\Delta}_{s.svy}^t \) is given by

\[ \text{var}(\hat{\Delta}_{s.svy}^t) = \sum_{k=1}^{K} \text{var}(\hat{\theta}_k), \] (2.43)

where

\[ \hat{\theta}_k = \frac{\hat{N}_{1k}}{\hat{N}_1} \left[ \hat{N}_{k1}^{-1} \sum_{i \in S} Z_i I(\hat{e}_i \in Q_k) Y_i \frac{1}{f(R_i)} - \hat{N}_{k0}^{-1} \sum_{i \in S} (1 - Z_i) I(\hat{e}_i \in Q_k) Y_i \frac{1}{f(R_i)} \right] . \]

The estimator \( \hat{\theta}_k \) can be written as a function of estimates of population totals

\[ \hat{\theta}_k = \left( \frac{\hat{U}_k}{\hat{N}_1} \right) \left( \frac{\hat{A}_k}{\hat{U}_k} - \frac{\hat{B}_k}{\hat{V}_k} \right) , \]

where \( \hat{U}_k = \hat{N}_{k1} \), \( \hat{V}_k = \hat{N}_{k0} \), \( \hat{A}_k = \sum_{i \in S} Z_i I(\hat{e}_i \in Q_k) Y_i \frac{1}{f(R_i)} \), and \( \hat{B}_k = \sum_{i \in S} (1 - Z_i) I(\hat{e}_i \in Q_k) Y_i \frac{1}{f(R_i)} \).

Just as before, we use a Taylor series linearization to approximate the variance.

The derivation is extremely similar to what we described earlier in Section 2.6.1, except now the partial derivatives of \( \hat{\theta}_k \) will be slightly different. All other aspects of the approach are the same.

### 2.7.2 Weighting Estimators

As we found when estimating \( \Delta_{PATE} \), multiplying the sampling weight and the propensity score weight leads us to estimators that estimate \( \Delta_{PATT} \). These estimators resemble equations (1.23), (1.25), and (1.29), which are the estimators in the non-survey situation. We estimate \( \Delta_{PATT} \) by

\[ \hat{\Delta}_{ipw1.svy}^t = \frac{1}{N_1} \sum_{i \in S} Z_i Y_i \frac{1}{f_i} - \frac{1}{N_1} \sum_{i \in S} (1 - Z_i) Y_i \hat{e}_i \frac{1}{f_i(1 - \hat{e}_i)} , \] (2.44)
where \( \hat{N}_1 = \sum_{i \in S} \frac{Z_i}{f_i} \). The estimator \( \hat{\Delta}_{ipw2.svy}^t \) is given by

\[
\hat{\Delta}_{ipw2.svy}^t = \left( \hat{N}_1 \right)^{-1} \sum_{i \in S} \frac{Z_i Y_i}{f_i} - \left( \sum_{i \in S} \frac{(1 - Z_i)\hat{e}_i}{f_i(1 - \hat{e}_i)} \right)^{-1} \sum_{i \in S} \frac{(1 - Z_i)Y_i\hat{e}_i}{f_i(1 - \hat{e}_i)}. \tag{2.45}
\]

This simply standardizes the weights compared to (2.44).

As before, these estimators belong to a more general class of estimators defined by the solutions to estimating equations

\[
\sum_{i \in S} \frac{1}{f_i} \left\{ Z_i (Y_i - \mu_1) \right\} = 0, \quad \text{and}
\sum_{i \in S} \frac{1}{f_i} \left\{ (1 - Z_i)(Y_i - \mu_0) e_i \right\} - \eta_0 \left( \frac{(Z_i - e_i)e_i}{1 - e_i} \right) = 0. \tag{2.46}
\]

Minimizing the variance of this class of estimators leads to the estimator

\[
\hat{\Delta}_{ipw3.svy}^t = \left( \hat{N}_1 \right)^{-1} \sum_{i \in S} \frac{1}{f_i} \left( Z_i Y_i \right) - \left\{ \sum_{i \in S} \frac{1}{f_i} \left( \frac{(1 - Z_i)\hat{e}_i}{1 - \hat{e}_i} \right) \left( 1 - \frac{C_0\hat{e}_i}{f_i(1 - \hat{e}_i)} \right) \right\}^{-1} \\
\times \sum_{i \in S} \frac{1}{f_i} \left( \frac{(1 - Z_i)Y_i\hat{e}_i}{(1 - \hat{e}_i)} \right) \left( 1 - \frac{C_0\hat{e}_i}{f_i(1 - \hat{e}_i)} \right) \\
= \hat{\mu}_{1,ipw3.svy}^t - \hat{\mu}_{0,ipw3.svy}^t, \text{ where}
\tag{2.47}
\]

\[
C_0 = -\sum_{i \in S} \frac{1}{f_i} \left( \frac{(Z_i - \hat{e}_i)\hat{e}_i}{1 - \hat{e}_i} \right) / \sum_{i \in S} \frac{1}{f_i^2} \left( \frac{(Z_i - \hat{e}_i)\hat{e}_i}{1 - \hat{e}_i} \right)^2.
\]

Estimators for the variance of \( \hat{\Delta}_{ipw1.svy}^t, \hat{\Delta}_{ipw2.svy}^t, \) and \( \hat{\Delta}_{ipw3.svy}^t \) are derived in a similar fashion to their corresponding estimators of the population average treatment effect in Section 2.6.2.

### 2.8 Simulation Study

We create a simulation to test the proposed estimators of \( \Delta_{PACE} \) in terms of bias, Monte Carlo variance, estimated variance, mean-squared error, and coverage. Under
several scenarios we simulate a population consisting of three covariates \((X_1, X_2, X_3)\), a pair of potential outcomes \((Y_1, Y_0)\), a treatment indicator \(Z\), and a sampling stratum variable \(R\). The true propensity score is determined by a known function of the covariates. We then take repeated samples of that population. The following steps are used to create the population:

1. Create four sampling strata of 10, 20, 30, and 40 thousand observations each.

2. Simulate \(X_3\), a Bernoulli random variable with \(p_1 = 0.1, p_2 = 0.3, p_3 = 0.5,\) and \(p_4 = 0.7\), where \(X_3 \sim Bernoulli(p_j)\) in the \(j\)th sampling strata.

3. Simulate \(X_1\) and \(X_2\). Conditional on \(X_3 = x_3\), \((X_1, X_2)' \sim N(\tau_{x_3}, \Sigma)\) where

\[
\tau_1 = (1, -1)', \tau_0 = (-1, 1)'
\]

and

\[
\Sigma = \begin{pmatrix} 1 & -0.5 \\ -0.5 & 1 \end{pmatrix}
\]

4. Simulate the true propensity score using the function

\[
e(X, \beta) = (1 + \exp(-(0 + \beta_1X_1 + \beta_2X_2 + \beta_3X_3)))^{-1}
\]

5. Simulate the treatment assignment \(Z\) using a Bernoulli random variable with probability equal to the true propensity score.

6. Simulate a pair of potential outcomes \((Y_1, Y_0)\) for each unit where

\[
Y_0 = 0 - 1X_1 + 1X_2 - 1X_3 + \epsilon, \quad \text{and}
\]

\[
Y_1 = Y_0 + 2 + \zeta X_1,
\]
where $\epsilon \sim N(0, \sigma = 1)$.

7. Create the observed outcome $Y = ZY_1 + (1 - Z)Y_0$.

We allow $\beta$ and $\zeta$ to vary to represent different degrees of association between $X$ and $Z$ as well as varying interaction effects. We use $\beta^{str} = (0.6, -0.6, 0.6)'$ and $\beta^{mod} = (0.3, -0.3, 0.3)'$, as well as $\zeta_0 = 0, \zeta_{sm} = 0.25, \zeta_{mod} = 0.5$, and $\zeta_{str} = 1$. Both potential outcomes are simulated in order to calculate the true population treatment effect.

We choose samples of size $n = 1,000$ from the population, with 250 observations chosen from each sampling strata. Sampling weights are calculated by $SW_j = N_j/250$ for all units in sampling strata $j$, where $N_j$ is the number of population units in sampling strata $j$. We estimate the propensity score both with and without utilizing the survey weights and create stratified and weighting estimators for each set of propensity scores.

We estimate the population treatment effect using the following estimators: the naive stratified estimator $\hat{\Delta}_s$, the survey-adjusted estimators $\hat{\Delta}_{s,svy1}$ and $\hat{\Delta}_{s,svy2}$, the naive weighting estimator $\hat{\Delta}_{ipw2}$, and the survey-adjusted weighting estimators $\hat{\Delta}_{ipw2,svy1}, \hat{\Delta}_{ipw2,svy2}$, the naive weighting estimator $\hat{\Delta}_{ipw3}$, and the survey adjusted estimator $\hat{\Delta}_{ipw3,svy}$. The estimator $\hat{\Delta}_{s,svy1}$ uses the unweighted propensity scores while $\hat{\Delta}_{s,svy2}$ uses the weighted propensity scores. Likewise $\hat{\Delta}_{ipw2,svy1}$ and $\hat{\Delta}_{ipw2,svy2}$ use the unweighted and weighted propensity scores. $\hat{\Delta}_{ipw3,svy}$ uses weighted propensity scores. In each case the correct form of the propensity score is used. All stratification estimators use five propensity score strata, and these are created using the quintiles of the propensity score. We also estimate the treatment effect using several regression estimators for reference. The regression models do not use the interaction term that is used in the true outcome model, but are otherwise correctly specified. These
include a regression estimator $\hat{\Delta}_{reg}$, a survey-weighted regression estimator $\hat{\Delta}_{reg.svy}$, and a survey-weighted regression estimator using the product of the propensity score weights and the survey weights $\hat{\Delta}_{reg.pssvy}$ as the weight. The estimator $\hat{\Delta}_{reg.pssvy}$ showed strong results in a previous simulation study [9].

We estimate the variance of weighting estimators using the formulas from earlier in the chapter. The survey-weighted stratification estimators use design-specific estimates of the variance under the assumptions in (2.36). The variance of the regression estimates are typical linear model or survey-weighted linear model estimates.

The simulation is implemented in R [43] with 10,000 samples taken for each scenario. MSE is calculated using the Monte Carlo (MC) variance along with the bias, and the MC coverage is calculated by creating 95% Wald confidence intervals from each estimate and estimated variance. Figure 2.3 shows the unweighted and weighted estimated propensity scores from one sample from a population with $\beta = \beta_{str}$. In this simulation there is very little difference between estimates of the propensity scores from the weighted and un-weighted models because they both have the correct parametric model; however, we do see differences between the estimators that use the weighted and unweighted propensity scores.

Figure 2.4 shows the percent bias of the estimators under the $\beta = \beta_{str}$ and $\beta = \beta_{mod}$ scenarios. In the following figures we only include one survey-adjusted propensity score stratification estimator $\hat{\Delta}_{s.svy}$. This is because there was no practical difference between using the weighted and unweighted propensity score. This was also the case with the regression estimator $\hat{\Delta}_{reg.pssvy}$. Full simulation results can be found in Table 2.1 and Table 2.2. In Figure 2.4 as the coefficient for the interaction, $\zeta$, increases, the bias of the naive estimators increases. When there is no interaction effect, all weighting and regression estimators are approximately unbiased. The
stratification estimators are initially biased as expected. As the interaction effect increases, so do the gains of using the survey weighted estimators compared to their naive counterparts. Also note that there are minimal differences between the bias of the weighting survey estimators when we estimate the propensity score with a weighted or unweighted logistic regression.

In Figure 2.5 we plot the mean-squared error of the estimators under the $\beta = \beta_{str}$ and $\beta = \beta_{mod}$ scenarios. While we are never able to get to the efficiency of $\hat{\Delta}_{reg.pssvy}$, the MSE of $\hat{\Delta}_{ipw3.svy}$ is very close. While stratification estimators may be more biased than weighting estimators, they do in general have smaller variances, as can be seen here by their often equivalent MSE. Among weighting estimators, $\hat{\Delta}_{ipw2.svy}$ substantially benefits from using the propensity scores estimated from weighted logistic regression rather than unweighted regression. The estimator $\hat{\Delta}_{ipw3.svy}$ is clearly the best choice in terms of MSE among all the weighting estimators and stratification estimators in this simulation.

Figure 2.6 shows the coverage of the 95% CI’s of the estimators as constructed from the estimated variance. Under the $\beta_{str}$ scenario, the coverage of each weighting estimator is slightly below its nominal level, but both reach the nominal level under the $\beta_{mod}$ scenario. This is because the variance estimates in the $\beta_{str}$ scenario are on average slightly less than the Monte Carlo simulation variance. This same pattern is seen in the Lunceford and Davidian simulations [27] from which our simulation was based. Stratification estimators are far below their nominal levels because the estimators are biased, not because of the performance of the proposed variance estimator.
2.9 Tables and Figures

Figure 2.1: Graphical representation of the general potential outcomes framework.

Figure 2.2: Graphical representation of a typical potential outcomes framework.
Figure 2.3: Estimated propensity scores using weighted and unweighted logistic regression from one iteration of the simulation study.

Figure 2.4: Percent bias of propensity score stratification, weighting, and regression estimators.
Figure 2.5: Mean squared error of propensity score stratification, weighting, and regression estimators.

Figure 2.6: Coverage of the 95% confidence intervals of the propensity score stratification and weighting estimators.
<table>
<thead>
<tr>
<th>ζ₀</th>
<th>Bias</th>
<th>Λ_{ipw2}</th>
<th>Λ_{ipw3}</th>
<th>Λ_{ipw2,svy}</th>
<th>Λ_{ipw3,svy}</th>
<th>Λ_s</th>
<th>Λ_{svy}</th>
<th>Λ_{reg}</th>
<th>Λ_{reg,svy}</th>
<th>Λ_{reg,psvy}</th>
<th>Λ_{reg,psvy}²</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bias</td>
<td>0.015</td>
<td>-0.027</td>
<td>-0.049</td>
<td>-0.021</td>
<td>-0.048</td>
<td>-0.275</td>
<td>-0.296</td>
<td>0.008</td>
<td>0.005</td>
<td>0.009</td>
<td>0.009</td>
</tr>
<tr>
<td>% Bias</td>
<td>0.768</td>
<td>-1.330</td>
<td>-2.465</td>
<td>-1.062</td>
<td>-2.408</td>
<td>-13.736</td>
<td>-14.778</td>
<td>0.391</td>
<td>0.236</td>
<td>0.447</td>
<td>0.443</td>
</tr>
<tr>
<td>MC Var</td>
<td>0.078</td>
<td>0.021</td>
<td>0.117</td>
<td>0.095</td>
<td>0.025</td>
<td>0.012</td>
<td>0.016</td>
<td>0.006</td>
<td>0.008</td>
<td>0.012</td>
<td>0.012</td>
</tr>
<tr>
<td>MSE</td>
<td>0.079</td>
<td>0.021</td>
<td>0.120</td>
<td>0.095</td>
<td>0.028</td>
<td>0.088</td>
<td>0.103</td>
<td>0.007</td>
<td>0.008</td>
<td>0.012</td>
<td>0.012</td>
</tr>
<tr>
<td>Coverage</td>
<td>0.907</td>
<td>0.896</td>
<td>0.827</td>
<td>0.884</td>
<td>0.872</td>
<td>0.293</td>
<td>0.328</td>
<td>0.333</td>
<td>0.950</td>
<td>0.946</td>
<td>0.935</td>
</tr>
<tr>
<td>ζ¹m</td>
<td>Bias</td>
<td>-0.013</td>
<td>-0.066</td>
<td>0.024</td>
<td>0.001</td>
<td>-0.040</td>
<td>-0.309</td>
<td>-0.285</td>
<td>-0.060</td>
<td>-0.039</td>
<td>0.005</td>
</tr>
<tr>
<td>% Bias</td>
<td>-0.675</td>
<td>-3.289</td>
<td>1.201</td>
<td>0.034</td>
<td>-2.007</td>
<td>-15.437</td>
<td>-14.228</td>
<td>-14.208</td>
<td>-2.987</td>
<td>-1.968</td>
<td>0.257</td>
</tr>
<tr>
<td>MC Var</td>
<td>0.082</td>
<td>0.021</td>
<td>0.124</td>
<td>0.095</td>
<td>0.025</td>
<td>0.013</td>
<td>0.016</td>
<td>0.007</td>
<td>0.008</td>
<td>0.013</td>
<td>0.013</td>
</tr>
<tr>
<td>MSE</td>
<td>0.082</td>
<td>0.025</td>
<td>0.125</td>
<td>0.095</td>
<td>0.026</td>
<td>0.108</td>
<td>0.097</td>
<td>0.010</td>
<td>0.010</td>
<td>0.013</td>
<td>0.013</td>
</tr>
<tr>
<td>Coverage</td>
<td>0.884</td>
<td>0.859</td>
<td>0.866</td>
<td>0.894</td>
<td>0.886</td>
<td>0.216</td>
<td>0.359</td>
<td>0.359</td>
<td>0.885</td>
<td>0.928</td>
<td>0.933</td>
</tr>
<tr>
<td>ζ²mod</td>
<td>Bias</td>
<td>-0.124</td>
<td>-0.147</td>
<td>-0.046</td>
<td>-0.032</td>
<td>-0.060</td>
<td>-0.358</td>
<td>-0.282</td>
<td>-0.281</td>
<td>-0.149</td>
<td>-0.098</td>
</tr>
<tr>
<td>MC Var</td>
<td>0.075</td>
<td>0.019</td>
<td>0.111</td>
<td>0.094</td>
<td>0.024</td>
<td>0.012</td>
<td>0.015</td>
<td>0.015</td>
<td>0.007</td>
<td>0.009</td>
<td>0.014</td>
</tr>
<tr>
<td>MSE</td>
<td>0.091</td>
<td>0.040</td>
<td>0.113</td>
<td>0.095</td>
<td>0.028</td>
<td>0.140</td>
<td>0.095</td>
<td>0.094</td>
<td>0.030</td>
<td>0.019</td>
<td>0.014</td>
</tr>
<tr>
<td>Coverage</td>
<td>0.785</td>
<td>0.702</td>
<td>0.829</td>
<td>0.874</td>
<td>0.859</td>
<td>0.358</td>
<td>0.352</td>
<td>0.573</td>
<td>0.824</td>
<td>0.932</td>
<td>0.932</td>
</tr>
<tr>
<td>ζ³str</td>
<td>Bias</td>
<td>-0.272</td>
<td>-0.265</td>
<td>-0.095</td>
<td>-0.078</td>
<td>-0.074</td>
<td>-0.438</td>
<td>-0.263</td>
<td>-0.262</td>
<td>-0.310</td>
<td>-0.203</td>
</tr>
<tr>
<td>MC Var</td>
<td>0.059</td>
<td>0.020</td>
<td>0.095</td>
<td>0.079</td>
<td>0.025</td>
<td>0.013</td>
<td>0.018</td>
<td>0.018</td>
<td>0.010</td>
<td>0.013</td>
<td>0.021</td>
</tr>
<tr>
<td>MSE</td>
<td>0.133</td>
<td>0.090</td>
<td>0.104</td>
<td>0.085</td>
<td>0.031</td>
<td>0.205</td>
<td>0.087</td>
<td>0.106</td>
<td>0.054</td>
<td>0.022</td>
<td>0.022</td>
</tr>
<tr>
<td>Coverage</td>
<td>0.590</td>
<td>0.430</td>
<td>0.785</td>
<td>0.831</td>
<td>0.843</td>
<td>0.422</td>
<td>0.423</td>
<td>0.105</td>
<td>0.568</td>
<td>0.928</td>
<td>0.930</td>
</tr>
</tbody>
</table>
Table 2.2: Weighting and stratification simulation results with $\beta = \beta_{mod}$.

<table>
<thead>
<tr>
<th>$\zeta$</th>
<th>$\hat{\Delta}_{ipw}^2$</th>
<th>$\hat{\Delta}_{ipw}^3$</th>
<th>$\hat{\Delta}_{ipw2.svy}^1$</th>
<th>$\hat{\Delta}_{ipw2.svy}^2$</th>
<th>$\hat{\Delta}_{ipw3.svy}$</th>
<th>$\hat{\Delta}_{svy}^1$</th>
<th>$\hat{\Delta}_{svy}^2$</th>
<th>$\hat{\Delta}_{reg}^1$</th>
<th>$\hat{\Delta}_{reg.svy}^1$</th>
<th>$\hat{\Delta}_{reg.pssvy}^1$</th>
<th>$\hat{\Delta}_{reg.pssvy}^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\zeta_0$</td>
<td>Bias</td>
<td>0.004</td>
<td>0.001</td>
<td>0.001</td>
<td>0.005</td>
<td>0.000</td>
<td>-0.154</td>
<td>-0.161</td>
<td>-0.161</td>
<td>0.004</td>
<td>0.003</td>
</tr>
<tr>
<td></td>
<td>% Bias</td>
<td>0.181</td>
<td>0.050</td>
<td>0.072</td>
<td>0.226</td>
<td>0.018</td>
<td>-7.697</td>
<td>-8.054</td>
<td>-8.051</td>
<td>0.186</td>
<td>0.143</td>
</tr>
<tr>
<td></td>
<td>MC Var</td>
<td>0.008</td>
<td>0.006</td>
<td>0.020</td>
<td>0.010</td>
<td>0.008</td>
<td>0.007</td>
<td>0.009</td>
<td>0.009</td>
<td>0.005</td>
<td>0.006</td>
</tr>
<tr>
<td></td>
<td>MSE</td>
<td>0.008</td>
<td>0.006</td>
<td>0.020</td>
<td>0.010</td>
<td>0.008</td>
<td>0.031</td>
<td>0.035</td>
<td>0.035</td>
<td>0.005</td>
<td>0.006</td>
</tr>
<tr>
<td></td>
<td>Avg Est Var</td>
<td>0.008</td>
<td>0.006</td>
<td>0.010</td>
<td>0.010</td>
<td>0.008</td>
<td>0.008</td>
<td>0.010</td>
<td>0.010</td>
<td>0.005</td>
<td>0.006</td>
</tr>
<tr>
<td></td>
<td>Coverage</td>
<td>0.951</td>
<td>0.950</td>
<td>0.953</td>
<td>0.950</td>
<td>0.579</td>
<td>0.620</td>
<td>0.628</td>
<td>0.951</td>
<td>0.948</td>
<td>0.947</td>
</tr>
<tr>
<td>$\zeta_{sm}$</td>
<td>Bias</td>
<td>-0.055</td>
<td>-0.058</td>
<td>0.016</td>
<td>0.001</td>
<td>-0.201</td>
<td>-0.150</td>
<td>-0.151</td>
<td>-0.124</td>
<td>-0.038</td>
<td>-0.002</td>
</tr>
<tr>
<td></td>
<td>% Bias</td>
<td>-2.752</td>
<td>-2.878</td>
<td>0.806</td>
<td>0.055</td>
<td>-10.048</td>
<td>-7.522</td>
<td>-7.527</td>
<td>-3.295</td>
<td>-0.720</td>
<td>0.205</td>
</tr>
<tr>
<td></td>
<td>MC Var</td>
<td>0.008</td>
<td>0.007</td>
<td>0.019</td>
<td>0.010</td>
<td>0.008</td>
<td>0.007</td>
<td>0.009</td>
<td>0.009</td>
<td>0.005</td>
<td>0.006</td>
</tr>
<tr>
<td></td>
<td>MSE</td>
<td>0.011</td>
<td>0.010</td>
<td>0.020</td>
<td>0.010</td>
<td>0.008</td>
<td>0.048</td>
<td>0.032</td>
<td>0.032</td>
<td>0.009</td>
<td>0.006</td>
</tr>
<tr>
<td></td>
<td>Avg Est Var</td>
<td>0.008</td>
<td>0.007</td>
<td>0.010</td>
<td>0.010</td>
<td>0.008</td>
<td>0.007</td>
<td>0.009</td>
<td>0.009</td>
<td>0.005</td>
<td>0.006</td>
</tr>
<tr>
<td></td>
<td>Coverage</td>
<td>0.890</td>
<td>0.880</td>
<td>0.837</td>
<td>0.951</td>
<td>0.947</td>
<td>0.355</td>
<td>0.655</td>
<td>0.656</td>
<td>0.842</td>
<td>0.946</td>
</tr>
<tr>
<td>$\zeta_{mod}$</td>
<td>Bias</td>
<td>-0.095</td>
<td>-0.099</td>
<td>-0.005</td>
<td>0.004</td>
<td>-0.001</td>
<td>-0.245</td>
<td>-0.152</td>
<td>-0.152</td>
<td>-0.124</td>
<td>-0.038</td>
</tr>
<tr>
<td></td>
<td>% Bias</td>
<td>-4.754</td>
<td>-4.971</td>
<td>-0.232</td>
<td>0.212</td>
<td>-12.265</td>
<td>-7.596</td>
<td>-7.585</td>
<td>-6.190</td>
<td>-1.913</td>
<td>-0.084</td>
</tr>
<tr>
<td></td>
<td>MC Var</td>
<td>0.009</td>
<td>0.007</td>
<td>0.019</td>
<td>0.011</td>
<td>0.008</td>
<td>0.007</td>
<td>0.009</td>
<td>0.009</td>
<td>0.005</td>
<td>0.007</td>
</tr>
<tr>
<td></td>
<td>MSE</td>
<td>0.018</td>
<td>0.017</td>
<td>0.019</td>
<td>0.011</td>
<td>0.008</td>
<td>0.067</td>
<td>0.032</td>
<td>0.032</td>
<td>0.020</td>
<td>0.008</td>
</tr>
<tr>
<td></td>
<td>Avg Est Var</td>
<td>0.009</td>
<td>0.007</td>
<td>0.011</td>
<td>0.011</td>
<td>0.009</td>
<td>0.007</td>
<td>0.009</td>
<td>0.009</td>
<td>0.005</td>
<td>0.007</td>
</tr>
<tr>
<td></td>
<td>Coverage</td>
<td>0.790</td>
<td>0.764</td>
<td>0.850</td>
<td>0.949</td>
<td>0.949</td>
<td>0.188</td>
<td>0.642</td>
<td>0.643</td>
<td>0.593</td>
<td>0.927</td>
</tr>
<tr>
<td>$\zeta_{str}$</td>
<td>Bias</td>
<td>-0.199</td>
<td>-0.200</td>
<td>0.017</td>
<td>-0.002</td>
<td>-0.323</td>
<td>-0.136</td>
<td>-0.137</td>
<td>-0.240</td>
<td>-0.074</td>
<td>-0.003</td>
</tr>
<tr>
<td></td>
<td>% Bias</td>
<td>-9.954</td>
<td>-10.024</td>
<td>0.860</td>
<td>-0.118</td>
<td>-0.169</td>
<td>-16.166</td>
<td>-6.820</td>
<td>-6.848</td>
<td>-11.995</td>
<td>-3.686</td>
</tr>
<tr>
<td></td>
<td>MC Var</td>
<td>0.010</td>
<td>0.008</td>
<td>0.019</td>
<td>0.013</td>
<td>0.010</td>
<td>0.009</td>
<td>0.011</td>
<td>0.011</td>
<td>0.007</td>
<td>0.009</td>
</tr>
<tr>
<td></td>
<td>MSE</td>
<td>0.049</td>
<td>0.049</td>
<td>0.020</td>
<td>0.013</td>
<td>0.010</td>
<td>0.113</td>
<td>0.030</td>
<td>0.030</td>
<td>0.065</td>
<td>0.015</td>
</tr>
<tr>
<td></td>
<td>Avg Est Var</td>
<td>0.010</td>
<td>0.008</td>
<td>0.012</td>
<td>0.012</td>
<td>0.007</td>
<td>0.009</td>
<td>0.009</td>
<td>0.007</td>
<td>0.009</td>
<td>0.010</td>
</tr>
<tr>
<td></td>
<td>Coverage</td>
<td>0.472</td>
<td>0.409</td>
<td>0.881</td>
<td>0.944</td>
<td>0.942</td>
<td>0.051</td>
<td>0.686</td>
<td>0.678</td>
<td>0.174</td>
<td>0.879</td>
</tr>
</tbody>
</table>
Chapter 3: Stratification-Weighting Hybrid Estimators

In general, stratified estimators are criticized for being biased because of possible residual confounding within each stratum; however, we have found that stratified estimators are sometimes able to produce smaller variances than inverse probability weighting estimators. Additionally, weighting estimators are sensitive to misspecification of the propensity score model and can be heavily influenced by large weights caused by extremely small or large propensity scores. To correct for residual confounding within strata which causes bias, researchers will sometimes use estimators such as \( \hat{\Delta}_{sr} \) (1.21), which uses a linear regression outcome model inside of each stratum. When the regression model is properly specified the estimator does very well. However, like all parametric models, it is subject to misspecification.

In this chapter we propose a propensity score estimator that takes on the attractive properties of both stratification and weighting estimators and does not rely on a parametric outcome model. The general idea is to estimate the treatment effect within a propensity score stratum with a weighting estimator. Within a propensity score stratum, the propensity scores and thus the propensity score weights will be similar, limiting the influence of large weights. We expect these estimators to be consistent and to have relatively small variances. We also think that these hybrid estimators will be more robust to model misspecification and extreme weights. We first develop these ideas in the non-survey context proposing point and variance estimators, and then
we extend them to a survey environment. Here we focus exclusively on estimators for $\Delta_{ATE}$ and $\Delta_{PATE}$, though these results can be easily extended to estimate other estimands. Results for simulations are shown and discussed.

### 3.1 Hybrid Estimators in the Non-Survey Context

We begin in the non-survey context where we assume our data is a simple random sample from the population of interest. We propose estimators that mimic $\hat{\Delta}_{ipw2}$ and $\hat{\Delta}_{ipw3}$ within each propensity score stratum. Specifically define

\[
\hat{\Delta}_{h.ipw2} = \frac{\sum_{k=1}^{K} n_k}{n} \left\{ \left( \sum_{i=1}^{n} \frac{Z_i I(\hat{e}_i \in Q_k)}{\hat{e}_i} \right)^{-1} \sum_{i=1}^{n} \frac{Z_i Y_i I(\hat{e}_i \in Q_k)}{\hat{e}_i} \right\} - \left( \sum_{i=1}^{n} \frac{(1 - Z_i) I(\hat{e}_i \in Q_k)}{(1 - \hat{e}_i)} \right)^{-1} \sum_{i=1}^{n} \frac{(1 - Z_i) Y_i I(\hat{e}_i \in Q_k)}{(1 - \hat{e}_i)} \right\}, \quad \text{and} \quad (3.1)
\]

\[
\hat{\Delta}_{h.ipw3} = \frac{\sum_{k=1}^{K} n_k}{n} \left\{ \left( \sum_{i=1}^{n} \frac{Z_i I(\hat{e}_i \in Q_k)}{\hat{e}_i} \left( 1 - \frac{C_{1k}}{\hat{e}_i} \right) \right)^{-1} \sum_{i=1}^{n} \frac{Z_i Y_i I(\hat{e}_i \in Q_k)}{\hat{e}_i} \left( 1 - \frac{C_{1k}}{\hat{e}_i} \right) \right\} - \left( \sum_{i=1}^{n} \frac{(1 - Z_i) I(\hat{e}_i \in Q_k)}{(1 - \hat{e}_i)} \left( 1 - \frac{C_{0k}}{1 - \hat{e}_i} \right) \right)^{-1} \times \sum_{i=1}^{n} \frac{(1 - Z_i) Y_i I(\hat{e}_i \in Q_k)}{(1 - \hat{e}_i)} \left( 1 - \frac{C_{0k}}{1 - \hat{e}_i} \right) \right\}, \quad (3.2)
\]

where

\[
C_{1k} = \frac{\sum_{i=1}^{n} I(\hat{e}_i \in Q_k) \{ (Z_i - \hat{e}_i) / \hat{e}_i \} \} } {\sum_{i=1}^{n} I(\hat{e}_i \in Q_k) \{ (Z_i - \hat{e}_i) / \hat{e}_i \}^2 } , \text{ and} \]

\[
C_{0k} = -\frac{\sum_{i=1}^{n} I(\hat{e}_i \in Q_k) \{ (Z_i - \hat{e}_i) / (1 - \hat{e}_i) \} \} } {\sum_{i=1}^{n} I(\hat{e}_i \in Q_k) \{ (Z_i - \hat{e}_i) / (1 - \hat{e}_i) \}^2 } ,
\]
where \( K \) is the number of propensity score strata, \( Q_k \) defines the \( k \)th stratum, and \( n_k \) is the number of observations in propensity score stratum \( k \). We estimate the variance of these estimators by

\[
\hat{I}_{h.ipw2,i} = \frac{Z_i(Y_i - \hat{\mu}_{1pss(i),h.ipw2})}{\hat{e}_i} - \frac{(1 - Z_i)(Y_i - \hat{\mu}_{0pss(i),h.ipw2})}{1 - \hat{e}_i} - (Z_i - \hat{e}_i)\hat{H}^{-1}_{\beta,2} \hat{E}_{\beta\beta} X_i, \tag{3.3}
\]

\[
\hat{I}_{h.ipw3,i} = \frac{Z_i(Y_i - \hat{\mu}_{1pss(i),h.ipw3}) + \hat{\eta}_{1pss(i)}(Z_i - \hat{e}_i)}{\hat{e}_i} - \frac{(1 - Z_i)(Y_i - \hat{\mu}_{0pss(i),h.ipw3}) - \hat{\eta}_{0pss(i)}(Z_i - \hat{e}_i)}{1 - \hat{e}_i} - (Z_i - \hat{e}_i)\hat{H}^{-1}_{\beta,3} \hat{E}_{\beta\beta} X_i, \tag{3.4}
\]

\[
\hat{E}_{\beta\beta}^{-1} = n^{-1} \sum_{i=1}^{n} \hat{e}_i(1 - \hat{e}_i)X_iX_i',
\]

\[
\hat{H}_{\beta,2} = n^{-1} \sum_{i=1}^{n} \left\{ \frac{Z_i(Y_i - \hat{\mu}_{1pss(i),h.ipw2})(1 - \hat{e}_i)}{\hat{e}_i} + \frac{(1 - Z_i)(Y_i - \hat{\mu}_{0pss(i),h.ipw2})\hat{e}_i}{1 - \hat{e}_i} \right\} X_i,
\]

\[
\hat{H}_{\beta,3} = n^{-1} \sum_{i=1}^{n} \left\{ \frac{Z_i(Y_i - \hat{\mu}_{1pss(i),h.ipw3} + \hat{\eta}_{1pss(i)})(1 - \hat{e}_i)}{\hat{e}_i} + \frac{(1 - Z_i)(Y_i - \hat{\mu}_{0pss(i),h.ipw3} + \hat{\eta}_{0pss(i)})\hat{e}_i}{1 - \hat{e}_i} \right\} X_i,
\]

\[
\hat{\eta}_k = -\frac{\sum_{i=1}^{n} I(\hat{e}_i \in Q_k)\{Z_i(Y_i - \hat{\mu}_{1k,ipw3})/\hat{e}_i^2\}}{\sum_{i=1}^{n} I(\hat{e}_i \in Q_k)\{(Z_i - \hat{e}_i)/\hat{e}_i\}^2}, \text{ and}
\]

80
\[
\hat{\eta}_{0k} = -\frac{\sum_{i=1}^{n} I(\hat{e}_i \in Q_k)\{(1 - Z_i)(Y_i - \hat{\mu}_{0k,ipw3})/(1 - \hat{e}_i)^2\}}{\sum_{i=1}^{n} I(\hat{e}_i \in Q_k)\{(Z_i - \hat{e}_i)/(1 - \hat{e}_i)^2\}}.
\]

The function \(pss(i)\) gives the propensity score stratum \(k = 1, \ldots, K\) of the \(i\)th observation. The estimators \(\hat{\mu}_{1k,ipw2}, \hat{\mu}_{0k,ipw2}, \hat{\mu}_{1k,ipw3}\), and \(\hat{\mu}_{0k,ipw3}\) give the average potential outcome in the treatment or control group in the \(k\)th propensity score stratum. The estimators \(\hat{\eta}_{1k}\) and \(\hat{\eta}_{0k}\) are propensity score stratum specific estimators of \(\eta_1\) and \(\eta_0\).

Similarly to the previous stratification estimators we recommend in practice increasing the number of the propensity score strata until the overall estimator converges or until the variance of the estimator starts to increase greatly. The type of outcome, continuous or binary, will play a role in determining the best number of strata to use.

**Proposition 3.** Assume there are \(k = 1, \ldots, K\), propensity score strata with fixed boundaries and that \(p_k\), the proportion of observations in propensity score stratum \(k\), is fixed. Also assume that the propensity scores are known, \((Y^{z=1}, Y^{z=0}) \perp \perp Z|e\), and that treated and control groups have common support with regards to their propensity score distributions. Under these assumptions \(\hat{\Delta}_{h,ipw2}\) is a consistent estimator for \(\Delta_{PATE}\).

**Proof.**

Within a propensity score stratum we have an unbiased estimator of the average potential outcomes. Observe that
\[
E \left[ \frac{ZY}{e(X)} \left| e(X) \in Q_k \right. \right] = E \left\{ E \left[ \frac{I(Z = 1)Y_{z=1}}{e(X)} \left| Y_{z=1}, X \right. \right] \left| e(X) \in Q_k \right. \right\} \\
= E \left\{ \frac{Y_{z=1}}{e(X)} E \left[ I(Z = 1) \left| Y_{z=1}, X \right. \right] \left| e(X) \in Q_k \right. \right\} \\
= E \left\{ \frac{Y_{z=1}}{e(X)} e(X) \left| e(X) \in Q_k \right. \right\} \\
= E \left[ \frac{Y_{z=1}}{e(X)} e(X) \left| e(X) \in Q_k \right. \right].
\] (3.6)

Similarly \( E \left[ \frac{(1-Z)Y}{1-e(X)} \left| e(X) \in Q_k \right. \right] = E \left[ Y_{z=0} \left| e(X) \in Q_k \right. \right] \). Then note that

\[
E \left[ \sum_{i=1}^{n} \frac{Z_iY_i I(e_i \in Q_k)}{e_i} \right] = E \left[ \sum_{i:e_i \in Q_k} \frac{Z_iY_i}{e_i} \right] = n_k E \left[ Y_{z=1} \left| e(X) \in Q_k \right. \right]
\]
as a result of (3.6). Similarly \( E \left[ \sum_{i=1}^{n} \frac{(1-Z_i)Y_i I(e_i \in Q_k)}{(1-e_i)} \right] = n_k E \left[ Y_{z=0} \left| e(X) \in Q_k \right. \right]. \)

This implies that

\[
\frac{1}{n} \sum_{i=1}^{n} \frac{(1-Z_i)Y_i I(e_i \in Q_k)}{e_i} \xrightarrow{p} p_k E \left[ Y_{z=0} \left| e(X) \in Q_k \right. \right], \text{ and} \\
\frac{1}{n} \sum_{i=1}^{n} \frac{Z_iY_i I(e_i \in Q_k)}{1-e_i} \xrightarrow{p} p_k E \left[ Y_{z=1} \left| e(X) \in Q_k \right. \right].
\]

While \( \left( \sum_{i=1}^{n} \frac{Z_i I(e_i \in Q_k)}{e_i} \right)^{-1} \) is not unbiased for \( n_k^{-1} \), \( \sum_{i=1}^{n} \frac{Z_i I(e_i \in Q_k)}{e_i} \) is unbiased for \( n_k \). Therefore

\[
\frac{1}{n_k} \sum_{i=1}^{n} \frac{Z_i I(e_i \in Q_k)}{e_i} \xrightarrow{p} 1, \text{ and} \\
\frac{1}{n_k} \sum_{i=1}^{n} \frac{(1-Z_i)I(e_i \in Q_k)}{1-e_i} \xrightarrow{p} 1.
\]

Putting all these parts together and as a consequence of Slutsky’s theorem and the weak law of large numbers
\[
\sum_{k=1}^{K} \frac{n_k}{n} \left\{ \left( \sum_{i=1}^{n} \frac{Z_i I(\hat{e}_i \in Q_k)}{\hat{e}_i} \right)^{-1} \sum_{i=1}^{n} \frac{Z_i Y_i I(\hat{e}_i \in Q_k)}{\hat{e}_i} \right. \\
- \left. \left( \sum_{i=1}^{n} \frac{(1 - Z_i) I(\hat{e}_i \in Q_k))}{(1 - \hat{e}_i)} \right)^{-1} \sum_{i=1}^{n} \frac{(1 - Z_i) Y_i I(\hat{e}_i \in Q_k)}{(1 - \hat{e}_i)} \right\}
\]

\[
= \sum_{k=1}^{K} \left\{ \left( \frac{1}{n_k} \sum_{i=1}^{n} \frac{Z_i I(\hat{e}_i \in Q_k)}{\hat{e}_i} \right)^{-1} \frac{1}{n} \sum_{i=1}^{n} \frac{Z_i Y_i I(\hat{e}_i \in Q_k)}{\hat{e}_i} \right. \\
- \left. \left( \frac{1}{n_k} \sum_{i=1}^{n} \frac{(1 - Z_i) I(\hat{e}_i \in Q_k))}{(1 - \hat{e}_i)} \right)^{-1} \frac{1}{n} \sum_{i=1}^{n} \frac{(1 - Z_i) Y_i I(\hat{e}_i \in Q_k)}{(1 - \hat{e}_i)} \right\}
\]

\[
\xrightarrow{n \to \infty} \sum_{k=1}^{K} \frac{p_k}{p} E \left[ Y_{z=1} - Y_{z=0} \bigg| e(X) \in Q_k \right]
\]

\[
= E \left[ Y_{z=1} - Y_{z=0} \right].
\]

Therefore \( \hat{\Delta}_{h.ipw2} \) is consistent for \( \Delta_{PATE} \).

Similarly we can show that \( \hat{\Delta}_{h.ipw3} \) is consistent for \( \Delta_{PATE} \) because each estimator within a propensity score stratum will be consistent for \( E \left[ Y_{z=1} - Y_{z=0} \big| e(X) \in Q_k \right] \).

### 3.1.1 Simulation

We create a simulation to test the proposed estimators in terms of bias, Monte Carlo (MC) variance, estimated variance, mean-squared error, and MC coverage. In addition we consider using different numbers of propensity score strata for each of the estimators. Under several scenarios, we take repeated samples of \( n = 1000 \) consisting of three covariates \((X_1, X_2, X_3)\), a pair of potential outcomes \((Y_1, Y_0)\), and a treatment indicator \(Z\). The true propensity score is determined by a known function of the covariates. The following steps are used to simulate the sample:

1. Simulate \( X_3 \), a Bernoulli random variable with \( p = 0.2 \), where

\[ X_3 \sim Bernoulli(p). \]
2. Simulate $X_1$ and $X_2$. Conditional on $X_3 = x_3$, $(X_1, X_2)^T \sim N(\tau_{x_3}, \Sigma)$ where

$$
\tau_1 = (1, -1)^T, \tau_0 = (-1, 1)^T
$$

and

$$
\Sigma = \begin{pmatrix} 1 & -0.5 \\ -0.5 & 1 \end{pmatrix}
$$

3. Simulate the true propensity score using the function

$$
e(X, \beta) = (1 + e^{-(0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3)})^{-1}.
$$

4. Simulate the treatment assignment $Z$ using a Bernoulli random variable with probability equal to the true propensity score.

5. Simulate a pair of potential outcomes $(Y_1, Y_0)$ for each unit where

$$
Y = 0 - 1X_1 + 1X_2 - 1X_3 + 2Z + \epsilon,
$$

and $\epsilon \sim N(0, \sigma = 1)$.

We allow $\beta$ to vary to represent different degrees of association between $X$ and $Z$. We use $\beta^{str} = (0.6, -0.6, 0.6)^T$ and $\beta^{mod} = (0.3, -0.3, 0.3)^T$. This simulation setup is identical to one of the scenarios that Lunceford and Davidian considered in their work on propensity score stratification and weighting estimators [27].

We estimate the treatment effect with the estimators $\hat{\Delta}_s$, $\hat{\Delta}_{ipw2}$, $\hat{\Delta}_{ipw3}$, $\hat{\Delta}_{h.ipw2}$, and $\hat{\Delta}_{h.ipw3}$. For the stratified estimator and the hybrid estimators, we also compare the estimates using 5, 6, 7, 8, 9, and 10 propensity score strata created by quantiles of the propensity score. The simulation is implemented using R [43] with 10,000 samples taken for each scenario. MSE is calculated using the MC variance along with the bias,
and the MC coverage is calculated by creating 95% Wald confidence intervals from each estimate and estimated variance. A small percentage of simulations (< 0.5%) were not included in some of the calculations because they were unable to produce an estimate, due to either no treated or no control observations falling within a propensity score stratum. This was more frequent when a large number of strata was used. This only happened in the $\beta^{str}$ scenario.

Figure 3.1 gives the percent bias of the estimators for both the $\beta^{str}$ and $\beta^{mod}$ scenarios. We see that the percent bias of the typical stratified estimator decreases steadily as the number of strata increases. Also note that all the hybrid estimators are approximately unbiased.

We plot the MSE of the estimators in Figure 3.2. Note that as with the bias, the MSE of the typical stratified estimator also decreases with the number of strata until it is very near the MSE of the other estimators. We observe some erratic behavior with $\hat{\Delta}_{h.ipw3}$ that here can be seen from how the MSE jumps around. This is caused by unstable estimates when there are too few treated and/or control units in a particular stratum. This happens infrequently in our simulation, but due to the large number of repetitions of the simulation, the estimator sometimes producing very large estimates in this situation, it makes a large impact on the MSE. The estimator $\hat{\Delta}_{h.ipw2}$ with 6 or more strata achieves slightly lower MSE for the $\beta^{str}$ scenario than $\hat{\Delta}_{ipw3}$, which previously was the best estimator with a non-parametric outcome model.

Figure 3.4 shows the coverage of the estimators. The coverage of hybrid estimators is comparable to the IPW estimators, and the coverage of the typical stratification estimator increases as the number of strata increases.
Instead of using a fixed number of propensity score strata, we create an additional simulation in which we begin with two propensity score strata and increase the number based on a stopping rule, such as might be used in practice. Based on a study of optimal propensity score stratification it is recommended to increase the number of strata used until the overall estimate remains relatively constant or until the estimated variance greatly increases [30]. Using these recommendations as a guideline we can avoid the erratic behavior of the estimators when the number of treated or control observations are too few within a stratum. Our general stopping rules are a) if the point estimate changed greatly and b) if the estimated variance increased substantially. We created three more specific stopping rules for each stratification estimator 1) no maximum number of strata and a minimum of two treated or control observations per stratum, 2) no maximum number of strata and a minimum of five treated or control observations per stratum, and 3) a maximum of ten strata and a minimum of five treated or control observations per stratum. In practice using a large number of strata might be prohibitive to checking the balance of the covariates within each stratum, but we want to observe if there is any improvement from not setting a maximum number of strata. Ten strata is still a realistic number for which to check balance.

Figure 3.3 shows the MSE of of the estimators using all three stopping rules. In addition, we show the MSE of weighting estimators and a regression estimator that is correctly specified for comparison. Observe that the hybrid estimators all perform slightly better than \( \hat{\Delta}_{ipw3} \) and are relatively close to achieving the efficiency of the regression estimator.

The coverage of the stratification and hybrid estimators is shown in Figure 3.5. The coverage of the hybrid estimators is best under stratification rule 3 and in general
for $\hat{\Delta}_{h, ipw}$, indicating that the large number of strata used in rules 1 and to some degree rule 2 have a negative impact on the estimation of variance. For comparison, the coverage of the estimators in Figure 3.5 is similar to the coverage of the corresponding estimators with ten strata in Figure 3.4.

Table 3.1 shows the mean number of propensity score stratum for each estimator and each rule in the simulation. The average number strata used for the $\beta_{mod}$ are quite large, and while using this many strata leads to smaller MSE, we do worry about its practicality. In practice it is important to check the balance of the covariates in each stratum, and doing this for 40 different strata would be tedious. The gains accomplished by using that many strata do not seem to justify the added amount of effort, though we do recommend using more than the standard five strata as the data allows.

We create one additional simulation to examine our hypothesis that the hybrid estimators should perform better than the weighting estimators when the propensity score model is misspecified. In this case we consider misspecification in terms of the functional form of the propensity score model, not misspecification due to unmeasured confounders. Unmeasured confounders appear to negatively affect the estimators equally. To investigate results under misspecification of the propensity score model we alter the simulation above slightly. The true propensity score model is changed to

$$e(X, \beta) = \left(1 + \exp(-0 + \beta_1(X_1)^{-1} + \beta_2(X_2)^2 + \beta_3X_3 + 0.2X_3X_2))\right)^{-1},$$

but we still estimate the propensity score with a logistic regression model with strictly linear terms. Note that the standardized differences after propensity score weighting are still “balanced” under rules of thumb for practical situations with the percent standardized difference after weighting of $X_1$ and $X_3$ both generally being under 5% and the percent standardized difference of $X_2$ being around 10% in the $\beta_{str}$ scenario.
In the $\beta_{mod}$ scenario the percent standardized differences after weighting were even smaller with $X_1$ and $X_3$ both generally being under 3% and $X_2$ being around 5%.

The results from these simulations are presented in Table 3.2. Even though the weighting estimators are reasonably balanced, they still give fairly biased estimates with decreased coverage of their 95% confidence intervals. In comparison, as we hypothesized, hybrid estimators are more robust to model misspecification. The stratification estimator and hybrid estimators all use a variable number of strata with stratification rule 3) from above. The stratification estimator also appears to be more robust than weighting estimators, and it performs almost as well as $\hat{\Delta}_{h.ipw2}$ in terms of MSE and even better in terms of coverage. The hybrid estimator $\hat{\Delta}_{h.ipw2}$ performs better than $\hat{\Delta}_{h.ipw3}$, further supporting the argument that the additional weighting adjustments made in $\hat{\Delta}_{h.ipw3}$ are unnecessary and quite possibly lead to poorer performance.

Overall, our results indicate that the proposed hybrid estimators, specifically $\hat{\Delta}_{h.ipw2}$ are a useful alternative to weighting and stratification estimators, taking on attractive properties of each. When the propensity score is properly specified, they give a consistent and approximately unbiased estimate of the treatment effect with as good or slightly better efficiency and coverage. Additionally, like the stratification estimator, they appear more robust than weighting estimators under misspecification of the propensity score model, which would seem to occur frequently in practice. We next consider hybrid estimators in the complex survey context.

### 3.2 Hybrid Estimators in the Survey Context

Now in the survey setting, we use estimators $\hat{\Delta}_{ipw2.svy}$ and $\hat{\Delta}_{ipw3.svy}$ within each propensity score stratum to estimate $\Delta_{PATE}$. Specifically define
\begin{equation}
\hat{\Delta}_{h.ipw2.svy} = \sum_{k=1}^{K} \frac{\hat{N}_k}{N} \left\{ \left( \sum_{i \in S} \frac{Z_i I(\hat{e}_i \in Q_k)}{f_i \hat{e}_i} \right)^{-1} \sum_{i \in S} \frac{Z_i Y_i(\hat{e}_i \in Q_k)}{f_i \hat{e}_i} \right. \\
- \left( \sum_{i \in S} \frac{(1 - Z_i) I(\hat{e}_i \in Q_k))}{f_i (1 - \hat{e}_i)} \right)^{-1} \sum_{i \in S} \frac{(1 - Z_i) Y_i(\hat{e}_i \in Q_k))}{f_i (1 - \hat{e}_i)} \right\}, \quad \text{and} \quad (3.7)
\end{equation}

\begin{equation}
\hat{\Delta}_{h.ipw3.svy} = \sum_{k=1}^{K} \frac{\hat{N}_k}{N} \left\{ (n_{1k})^{-1} \sum_{i \in S} \frac{Z_i Y_i(\hat{e}_i \in Q_k)}{f_i \hat{e}_i} \left(1 - \frac{C_{1k}}{f_i \hat{e}_i}\right) \right. \\
- \left( n_{0k} \right)^{-1} \sum_{i \in S} \frac{(1 - Z_i) Y_i(\hat{e}_i \in Q_k)}{f_i (1 - \hat{e}_i)} \left(1 - \frac{C_{0k}}{f_i (1 - \hat{e}_i)} \right) \right\}, \quad (3.8)
\end{equation}

where

\begin{align*}
n_{1k} &= \sum_{i \in S} \frac{Z_i I(\hat{e}_i \in Q_k)}{f_i \hat{e}_i} \left(1 - \frac{C_{1k}}{f_i \hat{e}_i}\right), \\
n_{0k} &= \sum_{i \in S} \frac{(1 - Z_i) I(\hat{e}_i \in Q_k)}{f_i (1 - \hat{e}_i)} \left(1 - \frac{C_{0k}}{f_i (1 - \hat{e}_i)} \right), \\
C_{1k} &= \frac{\sum_{i \in S} I(\hat{e}_i \in Q_k)\{(Z_i - \hat{e}_i)/f_i \hat{e}_i\}}{\sum_{i \in S} I(\hat{e}_i \in Q_k)\{(Z_i - \hat{e}_i)/f_i \hat{e}_i\}^2}, \quad \text{and} \\
C_{0k} &= -\frac{\sum_{i \in S} I(\hat{e}_i \in Q_k)\{(Z_i - \hat{e}_i)/f_i (1 - \hat{e}_i)\}}{\sum_{i \in S} I(\hat{e}_i \in Q_k)\{(Z_i - \hat{e}_i)/f_i (1 - \hat{e}_i)\}^2}.
\end{align*}

We estimate the variance covariance matrix \( V_\infty(\hat{\theta}_j - \theta^o) \), where \( j = 2 \) and \( j = 3 \) refer to \( \hat{\Delta}_{h.ipw2.svy} \) and \( \hat{\Delta}_{h.ipw3.svy} \) by

\begin{equation}
\hat{V}_\infty(\hat{\theta}_j - \theta^o) = \hat{A}_j^{-1} \hat{B}_j \hat{A}_j'^{-1}, \quad \text{where} \quad (3.9)
\end{equation}

\begin{equation}
\hat{A}_2^{-1} = \begin{bmatrix}
1 & 0 & -\hat{H}_{\beta,2,1}^{-1} \\
0 & 1 & \hat{H}_{\beta,2,0}^{-1} \\
0 & 0 & \hat{E}_{\beta}^{-1}
\end{bmatrix}.
\end{equation}
\[
\hat{A}_3^{-1} = \begin{bmatrix}
1 & 0 & 0 & 0 & -\hat{H}'_{\beta,3,1} \hat{E}_{\beta\beta}^{-1} \\
0 & 1 & 0 & 0 & \hat{H}'_{\beta,3,0} \hat{E}_{\beta\beta}^{-1} \\
0^* & 0 & 0^* & 0 & 0^* \\
0 & 0^* & 0 & 0^* & 0^* \\
0 & 0 & 0 & 0 & \hat{E}_{\beta\beta}^{-1}
\end{bmatrix}.
\]

Note that in the above matrix, \(0^*\) are not necessarily zero vectors, but they are unimportant to our ultimate variance calculation because we only need a subset of the variance matrix. We thus do not explicitly write them out for the sake of clarity.

Additional components of the variance are

\[
\hat{E}_{\beta\beta} = \hat{N}^{-1} \sum_{i \in S} \frac{1}{f_i} \hat{e}_i (1 - \hat{e}_i) \mathbf{X}_i \mathbf{X}_i', \quad \hat{N} = \sum_{i \in S} \frac{1}{f_i},
\]

\[
\hat{H}_{\beta,2,1} = \hat{N}^{-1} \sum_{i \in S} \frac{1}{f_i} \left( \frac{Z_i(Y_i - \hat{\mu}_{1\text{pss}(i),h,ipw.svy.2})(1 - \hat{e}_i)}{\hat{e}_i} \right) \mathbf{X}_i,
\]

\[
\hat{H}_{\beta,2,0} = \hat{N}^{-1} \sum_{i \in S} \frac{1}{f_i} \left( \frac{(1 - Z_i)(Y_i - \hat{\mu}_{1\text{pss}(i),h,ipw.svy.2})}{(1 - \hat{e}_i)} \hat{e}_i \right) \mathbf{X}_i,
\]

\[
\hat{H}_{\beta,3,1} = \hat{N}^{-1} \sum_{i \in S} \frac{1}{f_i} \left( \frac{Z_i(Y_i - \hat{\mu}_{1\text{pss}(i),h,ipw.svy.3} + \hat{\eta}_{1\text{svy.pss}(i)})}{\hat{e}_i} (1 - \hat{e}_i) \right) \mathbf{X}_i,
\]

\[
\hat{H}_{\beta,3,0} = \hat{N}^{-1} \sum_{i \in S} \frac{1}{f_i} \left( \frac{(1 - Z_i)(Y_i - \hat{\mu}_{0\text{pss}(i),h,ipw.svy.3} + \hat{\eta}_{0\text{svy.pss}(i)})}{(1 - \hat{e}_i)} \hat{e}_i \right) \mathbf{X}_i,
\]

\[
\hat{B}_j = \hat{N}^{-1} \Sigma_{ggj} + \hat{N}^{-2} \Sigma_{\text{tot}_j},
\]

\[
\Sigma_{ggj} = \hat{N}^{-1} \sum_{i \in S} \frac{1}{f_i} g_j(Y_i, \hat{\theta}) g_j(Y_i, \hat{\theta})',
\]

90
\( \mathbf{g}_2(\mathbf{Y}_i, \hat{\theta}) = \begin{pmatrix} Z_i \left( \frac{Y_i - \hat{\mu}_{1pss}(i), h.ipw.svy.2}{\hat{e}_i} \right) \\ (1 - Z_i) \left( \frac{Y_i - \hat{\mu}_{0pss}(i), h.ipw.svy.2}{1 - \hat{e}_i} - \hat{e}_i \right) \mathbf{X}_i' \end{pmatrix}, \)

\( \mathbf{g}_3(\mathbf{Y}_i, \hat{\theta}) = \begin{pmatrix} \frac{Z_i(Y_i - \hat{\mu}_{1pss}(i), h.ipw.svy.3)}{\hat{e}_i} + \hat{\eta}_{1,svy.pss(i)} \left( \frac{Z_i - \hat{e}_i}{\hat{e}_i} \right) \\ \frac{(1 - Z_i)(Y_i - \hat{\mu}_{0pss}(i), h.ipw.svy.3)}{1 - \hat{e}_i} - \hat{\eta}_{0,svy.pss(i)} \left( \frac{Z_i - \hat{e}_i}{1 - \hat{e}_i} \right) \\ \hat{\eta}_{1,svy.pss(i)} \left( \frac{Z_i - \hat{e}_i}{\hat{e}_i} \right)^2 + \frac{Z_i(Y_i - \hat{\mu}_{1pss}(i), h.ipw.svy.3)}{\hat{e}_i^2} \\ \hat{\eta}_{0,svy.pss(i)} \left( \frac{Z_i - \hat{e}_i}{1 - \hat{e}_i} \right)^2 + \frac{(1 - Z_i)(Y_i - \hat{\mu}_{0pss}(i), h.ipw.svy.3)}{(1 - \hat{e}_i)^2} \left( Z_i - \hat{e}_i \right) \mathbf{X}_i' \end{pmatrix} \), and

\[ \Sigma_{tot_j} = \hat{V} \left( \sum_{i \in S} \frac{1}{f_i} \mathbf{g}_j(\mathbf{Y}_i, \hat{\theta}) \bigg| \mathcal{F}_N \right). \]

\( \Sigma_{tot_j} \) is the estimated sampling variance matrix of the survey total of \( \mathbf{g}_j(\mathbf{Y}_i, \hat{\theta}) \). This variance can be estimated by a design-specific estimator using standard statistical software.

For inference about \( \Delta_{PATE} \), we only care about a portion of \( \hat{V}_\infty(\hat{\theta}_j - \theta^o) \). The estimated variances of \( \hat{\Delta}_{h.ipw2.svy} \) and \( \hat{\Delta}_{h.ipw3.svy} \) can be computed by

\[ \hat{V}(\hat{\Delta}_{h.ipw2.svy}) = \begin{bmatrix} 1 & -1 & 0 \end{bmatrix} \hat{\mathbf{A}}_2^{-1} \hat{\mathbf{B}}_2 \hat{\mathbf{A}}^{-1}_2 \begin{bmatrix} 1 \\ -1 \\ 0 \end{bmatrix}, \text{ and} \]

\[ \hat{V}(\hat{\Delta}_{h.ipw3.svy}) = \begin{bmatrix} 1 & -1 & 0 \end{bmatrix} \hat{\mathbf{A}}_3^{-1} \hat{\mathbf{B}}_3 \hat{\mathbf{A}}^{-1}_3 \begin{bmatrix} 1 \\ -1 \\ 0 \end{bmatrix}. \]

### 3.2.1 Evaluating Balance

We evaluate balance after stratification and weighting for the hybrid estimators much the same way as for the stratified and weighted estimators. Checking the balance should be an intermediate step before we estimate the treatment effect. Like stratified estimators, the after-balance should be checked in each of the propensity
score strata. Instead of using equations (2.30) and (2.31) with weights $w_i = 1/f_i$ applied to the observations in each propensity score stratum, now apply weights based on the propensity score and probability of sampling. For $\hat{\Delta}_{h.ipw2.svy}$ use weights equal to

$$w_i = \frac{z_i}{f_i \hat{e}_i} + \frac{(1 - z_i)}{f_i (1 - \hat{e}_i)},$$

and for $\hat{\Delta}_{h.ipw3.svy}$ use

$$w_i = \frac{z_i}{f_i \hat{e}_i} \left( 1 - \frac{C_1(i)}{f_i \hat{e}_i} \right) + \frac{(1 - z_i)}{f_i (1 - \hat{e}_i)} \left( 1 - \frac{C_0(i)}{f_i (1 - \hat{e}_i)} \right),$$

where $C_1(i)$ and $C_0(i)$ are estimates of $C_1$ and $C_0$ from the propensity score stratum from which observation $i$ belongs.

### 3.2.2 Simulation

We investigate the performance of these hybrid estimators using the same simulation setup as in Section 2.8. As we did in Section 3.1.1, we use an increasing number of strata (5 through 10) to estimate the treatment effect. The results shown are only for the $\gamma = \gamma_{str}$ scenario, which uses a strong heterogeneous treatment effect.

Both hybrid estimators are approximately unbiased, while the bias of the stratification estimator decreases with the increased number of propensity score strata. Figure 3.6 shows the MSE of the different estimators under increasing numbers of propensity score strata. The MSE of $\hat{\Delta}_{h.ipw2.svy}$ is at or below that of $\hat{\Delta}_{ipw3.svy}$ when using at least nine propensity score strata. Meanwhile we see the same erratic behavior of $\hat{\Delta}_{h.ipw3.svy}$ as we did with $\hat{\Delta}_{h.ipw3}$ earlier in the chapter.
Figure 3.7 shows the coverage of the estimators. The hybrid estimators do about as well as the weighting estimators in the $\beta_{str}$ scenario, but provide slightly less coverage than the weighting estimators in the $\beta_{mod}$ scenario.

When using hybrid estimators we recommend the same advice as for previous stratification estimators. Create propensity score strata based on the quantiles of the propensity score and increase the number of strata until the point estimate seems to converge or the estimated variance increases dramatically. Just as in the non-survey setting, these hybrid estimators should provide some added robustness under misspecification of the propensity score compared to weighting estimators without sacrificing near-unbiasedness or coverage. Similarly to the non-survey setting, we recommend the use of $\hat{\Delta}_{h.ipu2.svy}$ over $\hat{\Delta}_{h.ipu3.svy}$ because of $\hat{\Delta}_{h.ipu3.svy}$'s sometimes erratic behavior.
3.3 Tables and Figures

Figure 3.1: Percent bias of propensity score stratification, weighting, and hybrid estimators with variable number of propensity score strata in a non-survey context.
Figure 3.2: MSE of propensity score stratification, weighting, and hybrid estimators with variable number of propensity score strata in a non-survey context.

Figure 3.3: MSE of propensity score stratification, weighting, hybrid, and regression estimators with stratification rules in a non-survey context.
Figure 3.4: Coverage of 95% confidence intervals created by propensity score strati-
fication, weighting, and hybrid estimators with variable number of propensity score
strata in a non-survey context.

Figure 3.5: Coverage of 95% confidence intervals created by propensity score strati-
fication and hybrid estimators with stratification rules in a non-survey context.
Table 3.1: Mean number of propensity score strata used in propensity score stratification and hybrid estimators with stratification rules in a non-survey context.

<table>
<thead>
<tr>
<th>Sim</th>
<th>$\hat{\Delta}_s^1$</th>
<th>$\hat{\Delta}_s^2$</th>
<th>$\hat{\Delta}_s^3$</th>
<th>$\hat{\Delta}_{h.ipw2}^1$</th>
<th>$\hat{\Delta}_{h.ipw2}^2$</th>
<th>$\hat{\Delta}_{h.ipw2}^3$</th>
<th>$\hat{\Delta}_{h.ipw3}^1$</th>
<th>$\hat{\Delta}_{h.ipw3}^2$</th>
<th>$\hat{\Delta}_{h.ipw3}^3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta_{str}$</td>
<td>17.2</td>
<td>10.0</td>
<td>9.1</td>
<td>16.8</td>
<td>10.0</td>
<td>9.1</td>
<td>16.1</td>
<td>10.0</td>
<td>9.1</td>
</tr>
<tr>
<td>$\beta_{mod}$</td>
<td>42.3</td>
<td>24.8</td>
<td>10</td>
<td>42.3</td>
<td>24.8</td>
<td>10.0</td>
<td>38.8</td>
<td>24.5</td>
<td>10.0</td>
</tr>
</tbody>
</table>

Table 3.2: Simulation results for propensity score stratification, weighting, and hybrid estimators when the propensity score model is misspecified in a non-survey context.

<table>
<thead>
<tr>
<th>Statistic</th>
<th>$\beta$</th>
<th>$\hat{\Delta}_s$</th>
<th>$\hat{\Delta}_{ipw2}$</th>
<th>$\hat{\Delta}_{ipw3}$</th>
<th>$\hat{\Delta}_{h.ipw2}$</th>
<th>$\hat{\Delta}_{h.ipw3}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>% Bias</td>
<td>$\beta_{str}$</td>
<td>-2.534</td>
<td>-9.626</td>
<td>-9.774</td>
<td>-0.092</td>
<td>-3.061</td>
</tr>
<tr>
<td>MC Variance</td>
<td>$\beta_{str}$</td>
<td>0.011</td>
<td>0.009</td>
<td>0.009</td>
<td>0.012</td>
<td>0.020</td>
</tr>
<tr>
<td>95% Coverage</td>
<td>$\beta_{str}$</td>
<td>0.903</td>
<td>0.451</td>
<td>0.398</td>
<td>0.851</td>
<td>0.711</td>
</tr>
<tr>
<td>MSE</td>
<td>$\beta_{str}$</td>
<td>0.014</td>
<td>0.046</td>
<td>0.047</td>
<td>0.012</td>
<td>0.024</td>
</tr>
<tr>
<td>% Bias</td>
<td>$\beta_{mod}$</td>
<td>-1.607</td>
<td>-4.022</td>
<td>-4.001</td>
<td>-0.275</td>
<td>-1.493</td>
</tr>
<tr>
<td>MC Variance</td>
<td>$\beta_{mod}$</td>
<td>0.007</td>
<td>0.005</td>
<td>0.005</td>
<td>0.007</td>
<td>0.008</td>
</tr>
<tr>
<td>95% Coverage</td>
<td>$\beta_{mod}$</td>
<td>0.966</td>
<td>0.802</td>
<td>0.800</td>
<td>0.909</td>
<td>0.879</td>
</tr>
<tr>
<td>MSE</td>
<td>$\beta_{mod}$</td>
<td>0.008</td>
<td>0.012</td>
<td>0.012</td>
<td>0.007</td>
<td>0.009</td>
</tr>
</tbody>
</table>
Figure 3.6: MSE of propensity score stratification, weighting, and hybrid estimators with variable number of propensity score strata in survey context with $\gamma = \gamma_{str}$.

Figure 3.7: Coverage of 95% confidence intervals created by propensity score stratification, weighting, and hybrid estimators with variable number of propensity score strata in a survey context with $\gamma = \gamma_{str}$. 
Chapter 4: Matching Estimators

Matching estimators are some of the most popular propensity score methods with practitioners because they are intuitive and relatively simple to understand. In this chapter we propose matching estimators to be used with complex survey data that do not assume a homogeneous treatment effect. As with the stratification and weighting estimators, we continue to only consider estimators with non-parametric outcome models. For example, we do not consider any estimators that utilize a regression model for the outcome after matching. We first propose the estimators and discuss their properties. Next we propose and describe estimators for the variance of the point estimators, and finally we conduct a simulation to further examine the properties of the estimators. Assume throughout this chapter that 1) for 1:K matching estimators the support in terms of the propensity score distribution of control observations is at least that of treated observations and 2) for full matching estimators the treated and control observations share the same support.

4.1 Point Estimators

We consider point estimators in the survey context that are analogs of 1:K and full matching estimators. While 1:K estimators are extremely popular in practice, we will explain why they might not be a good fit in many survey situations. In contrast, we
will explain how full matching estimators, which can be more difficult to implement, are adaptable to a variety of situations.

### 4.1.1 1:K Survey Matching Estimator

1:K matching is the most common matching estimator because it is intuitive and easy to implement using nearest neighbor matching. Assume we create $n_1$ matched sets matched on $\hat{e}$, each with a treated observation and $K$ control observations. Let $j = 1, \ldots, n_1$ index the matched sets, so then $Y_{j1}$ denotes the outcome for the treated observation in the $j$th matched set and $Y_{j0(1)}, \ldots, Y_{j0(K)}$ denote the outcomes for the $K$ control observations in the $j$th matched set. Then

$$\hat{\Delta}_{\text{matchK.svy}}^t = 1 \hat{N}_1 \sum_{j=1}^{n_1} \frac{1}{f_{j1}} (Y_{j1} - \hat{Y}_{j0})$$  \hspace{1cm} (4.1)

where $\hat{Y}_{j0} = (\sum_{k=1}^{K} \frac{1}{f_{j0(k)}})^{-1} \sum_{k=1}^{K} \frac{Y_{j0(k)}}{f_{j0(k)}}$ and $\hat{N}_1 = \sum_{j=1}^{n_1} \frac{1}{f_{j1}}$ is an estimator of $\Delta_{PATT}$. This estimator can also be written as

$$\hat{\Delta}_{\text{matchK.svy}}^t = 1 \hat{N}_1 \sum_{i=1}^{n} \frac{1}{f_i} Y_i \left( Z_i - (1 - Z_i) \left( \frac{W^T_i}{W^C_i} \right) \right)$$  \hspace{1cm} (4.2)

where $n$ is the number of treated and control observations in the matched sample and $W^T_i$ and $W^C_i$ are the sum of the treated and control survey weight in the matched set from which observation $i$ belongs, respectively.

The motivation for this estimator follows from the empirical mass function. First note that

$$\hat{f}(e_i) = \sum_{j=1}^{n_1} \frac{1}{f_{j1}} I(e_{j1} = e_i) \frac{1}{\hat{N}_1}$$

is the empirical mass function of the distribution of $e|Z = 1$ assuming that each treated propensity score has a unique value. Generally speaking the idea of $\hat{\Delta}_{\text{matchK.svy}}^t$
is to average the difference between $Y^{z=1}|e$ and $Y^{z=0}|e$ with respect to the distribution of $e|Z = 1$. This is the same idea used in the stratification estimator.

While the point estimator is relatively straightforward to compute, the process of creating $K$ matches for each and every treated unit may be difficult in a survey setting based on the distribution of the propensity score. Rosenbaum and Rubin [36] motivate their matching estimator by considering a two-step sampling process in which first a propensity score is randomly sampled, and then a treated unit with that propensity score is selected at random from the treated population and a control unit with that propensity score is selected at random from the control population. This essentially assumes that we have a sample from the treated population and a sample from the control population. In many applications this is actually the case. For example, we may find a treated sample of observations from, say, medical records, and then we gather a much larger suitable sample of controls. In some settings researchers may even be in control of the number of control units to sample.

Consider the situation where the distribution of the true propensity score in both the treated and control populations includes a value of $e_i = 0.75$. If we were to take a random sample of the entire population we would expect to sample 3 treated units with $e_i = 0.75$ for every control unit with that propensity score. This creates a problem for matching all the sampled units with $e_i = 0.75$. While we may be able to find a good match for some of the units, it becomes clear that we will not be able to find good matches for all of them. This is true even when the number of control units is many times larger than the number of treated units. By definition of the propensity score, we will not be able to find good 1:1 matches for all the treated units with a propensity score greater than a half under a simple random sample, and if we consider 1:2, 1:3, or greater matching the problem is compounded.
In contrast if we are sampling separately from treated and control populations as in the motivation of Rosenbaum and Rubin above, then it is possible to gather a large enough sample of controls to create good matches for all treated units, even those with large propensity scores. The key is that we are able to increase the sample size of control units without increasing the sample size of treated units with which we need to find matches.

When using a complex sample survey that does not sample control units at a much higher rate than treated units, we will be faced with a situation similar to that of taking a random sample of the population. We will not be able to find good matches for all treated units with propensity scores greater than a half (and sometimes less depending on the value of K) using 1:K matching. While this matching estimator works well in the right situations, it is not flexible enough to be able to deal with many situations for which we may want to use a matching estimator.

4.1.2 Survey Full Matching Estimator

While not as popular in practice as 1:K matching, full matching provides a flexible approach to matching and allows us to estimate either $\Delta_{PATE}$ or $\Delta_{PATT}$. Assume that we create $J$ matched sets, matching on $\hat{e}$, which consist of exactly one treated observation and at least one control observation, or exactly one control observation and at least one treated observation. It is related to stratification in the sense that full matching is the finest possible stratification [35]. We estimate $\Delta_{PATE}$ by

$$\hat{\Delta}_{match.full.svy} = \frac{1}{N} \sum_{i \in S} \frac{1}{f_i} \left( \hat{Y}_{i1} - \hat{Y}_{i0} \right), \quad (4.3)$$

where the data are divided into $j = 1, \ldots, J$ mutually exclusive sets with similar propensity scores, $\hat{Y}_{i1} = \hat{N}_{j(i)1}^{-1} \sum_{k=1}^{n_{j(i)}} \frac{Z_{j(i)k}Y_{j(i)k}}{f_{j(i)k}}$ and $\hat{Y}_{i0} = \hat{N}_{j(i)0}^{-1} \sum_{k=1}^{n_{j(i)}} \frac{(1-Z_{j(i)k})Y_{j(i)k}}{f_{j(i)k}}$ are
the average of the treated and control observations respectively from which observation \( i \) belongs. The function \( j(i) \) is equal to the matched set from which observation \( i \) belongs. Additionally, \( \hat{N}_{j(i)}^{(1)} = \sum_{k=1}^{n_{j(i)}} \frac{Z_{j(i)k}}{f_{j(i)k}} \) and \( \hat{N}_{j(i)}^{(0)} = \sum_{k=1}^{n_{j(i)}} \frac{1-Z_{j(i)k}}{f_{j(i)k}} \). We can also write \( \hat{\Delta}_{\text{matchfull.svy}} \) as

\[
\hat{\Delta}_{\text{matchfull.svy}} = \frac{1}{N} \sum_{i \in S} \frac{1}{f_i} Y_i \left( Z_i \left( \frac{W_i^A}{W_i^C} \right) - (1 - Z_i) \left( \frac{W_i^A}{W_i^C} \right) \right),
\]

(4.4)

where \( W_i^T \) is the sum of the treated survey weight in the matched set from which observation \( i \) belongs, and \( W_i^A \) is the sum of the treated and control survey weight in the matched set from which observation \( i \) belongs.

We can similarly estimate \( \Delta_{PATT} \) by

\[
\hat{\Delta}_{\text{matchfull.svy}}^t = \frac{1}{N_1} \sum_{i \in S} \frac{Z_i}{f_i} \left( \hat{Y}_{i1} - \hat{Y}_{i0} \right),
\]

(4.5)

where \( \hat{Y}_{i1} \) and \( \hat{Y}_{i0} \) are defined as above. This can also be written as

\[
\hat{\Delta}_{\text{matchfull.svy}}^t = \frac{1}{N_1} \sum_{i \in S} \frac{1}{f_i} Y_i \left( Z_i - (1 - Z_i) \left( \frac{W_i^T}{W_i^C} \right) \right),
\]

(4.6)

where \( W_i^T \) and \( W_i^C \) are the sum of the treated and control survey weight in the matched set from which observation \( i \) belongs, respectively.

The motivation for the full matching estimator is much the same as for the 1:K matching estimator. We average estimates of \( Y^{z=1}\lvert e \) and \( Y^{z=0}\lvert e \) with respect to the distribution of \( e \) (for \( \Delta_{PATE} \)) or \( e\lvert Z = 1 \) (for \( \Delta_{PATT} \)), and then we estimate the distribution of \( e \) and \( e\lvert Z = 1 \) using their empirical mass functions.

Full matching is extremely flexible because we are not setting a fixed ratio of treated to control units for a matched set. Thinking about the distribution of the propensity score, in areas where there are many more controls than treated (\( e < 0.25 \))
we can match multiple controls to a treated observation. In areas where there are many more treated observations than control \((e > 0.75)\) we can match multiple treated observations to a control observation. Utilizing all of the observations allows us to estimate \(\Delta_{PATE}\) as well as \(\Delta_{PATT}\).

### 4.2 Decomposition of the Estimators

Following Abadie and Imbens [1], we decompose our matching estimators into three parts, which will motivate discussions of bias and how we estimate the variance. First assume strongly ignorable treatment assignment given covariates \(X\). Then define

\[
\Delta(x) = E[Y_{z=1} - Y_{z=0} | X = x]
\]

as the conditional treatment effect given covariates \(x\). Also assume that the data generating process in the population is as follows

\[
Y_i = Z_i \mu_1(X_i) + (1 - Z_i) \mu_0(X_i) + \epsilon_i,
\]

where \(\mu_z(X_i) = E[Y | Z = z, X = x]\) and \(\epsilon_i\) is a randomly distributed variable with mean 0 and variance \(\sigma^2(X_i, Z_i)\).

Let \(\hat{\Delta}_M\) be one of our matching estimators in general. Then the difference between \(\hat{\Delta}_M\) and \(\Delta\) can be written as

\[
\hat{\Delta}_M - \Delta = \left(\overline{\Delta(X)} - \Delta\right) + E_M + B_M,
\]

(4.7)

where \(\overline{\Delta(X)}\) is the weighted average of the conditional treatment effect for our sample,

\[
\overline{\Delta(X)} = \frac{1}{N} \sum_{i \in S} \frac{1}{f_i} (\mu_1(X_i) - \mu_0(X_i)).
\]

(4.8)

\(E_M\) is the weighted average of the residuals, and \(B_M\) is the conditional bias relative to \(\overline{\Delta(X)}\) due to inexact matching. For example, take \(\hat{\Delta}_M = \hat{\Delta}_{matchfull.svy}\), then
\[ E_M = \frac{1}{N} \sum_{i \in S} \frac{1}{f_i} \left( Z_i \left( \frac{W_i^A}{W_i^T} \right) - (1 - Z_i) \left( \frac{W_i^A}{W_i^C} \right) \right) \epsilon_i, \quad \text{and} \quad (4.9) \]

\[
B_M = \frac{1}{N} \sum_{i \in S} \frac{1}{f_i} \left( Z_i \left( \frac{W_i^A}{W_i^T} \right) \sum_{k=1}^{n_j(i)} \frac{1/f_{j(i)k}}{W_i^C} (1 - Z_{j(i)k}) \left[ \mu_1(X_i) - \mu_1(X_{j(i)k}) \right] \right.
- \left. (1 - Z_i) \left( \frac{W_i^A}{W_i^C} \right) \sum_{k=1}^{n_j(i)} \frac{1/f_{j(i)k}}{W_i^T} (Z_{j(i)k}) \left[ \mu_0(X_i) - \mu_0(X_{j(i)k}) \right] \right).
\quad (4.10)\]

\(B_M\) is the weighted difference between the average conditional means for observations in a matched set. If matching is exact on \(X\), then \(B_M = 0\). In general this is not the case. Note that this decomposition is similar for \(\Delta = \Delta_{PATE}\) or \(\Delta = \Delta_{PATT}\).

When \(\Delta = \Delta_{PATT}\), replace \(\Delta(X)\) with \(\Delta(X)^t = \frac{1}{N_1} \sum_{i \in S} Z_{i}^t (\mu_1(X_i) - \mu_0(X_i))\). \(E_M\) and \(B_M\) will also be different for \(\hat{\Delta}_{\text{matchfull.svy}}^t\) and \(\hat{\Delta}_{\text{matchK.svy}}^t\).

### 4.3 Bias of the Estimators

**Proposition 4.** Assuming that we match exactly on \(X\) or that average bias across samples, \(E[B_M]\) is zero, then our estimators will be unbiased for \(\Delta_{PATE}\) or \(\Delta_{PATT}\) respectively.

**Proof.**

The bias of the matching estimators can be written using the decomposition as

\[
E \left[ \hat{\Delta}_M - \Delta \right] = E \left[ \left( \Delta(X) - \Delta \right) + E_M + B_M \right].
\]

As a result of our assumptions about the data generating process, \(E[\epsilon_i] = 0\), which implies that \(E[E_M] = 0\). The expected value of \(\Delta(X)\) is
\[ E[\Delta(X)] = E \left[ \frac{1}{N} \sum_{i=1}^{N} \frac{S_i}{f_i} (\mu_1(X_i) - \mu_0(X_i)) \right] \]
\[ = \frac{1}{N} \sum_{i=1}^{N} (\mu_1(X_i) - \mu_0(X_i)) \]
\[ = \Delta_{PATE} \]

assuming that the sampling weights are correctly specified. Likewise \( E[\Delta(X)^t] = \Delta_{PATT} \).

Then the only bias in our estimators comes from the bias due to inexact matching \( B_M \); however, we are assuming that we are able to match exactly on \( X \) or that average bias across samples, \( E[B_M] \) is zero. Therefore our estimators are unbiased.

For further discussion of the asymptotic properties of the matching discrepancy see Abadie and Imbens [1].

4.4 Variance Estimators

We use the decomposition of the estimators to motivate our estimates of variance. Ignoring \( B_M \) and assuming the weighted average of the residuals \( E_M \) is independent of \( \Delta(X) \), the variance of \( \hat{\Delta}_M \) can be broken down into two components:

\[ V(\hat{\Delta}_M) = V(\bar{\Delta}(X)) + V(E_M) \]

We estimate \( V(\bar{\Delta}(X)) \) and \( V(\bar{\Delta}(X)^t) \) by

\[ \hat{V}_1 = \hat{V}_{svy} \left( \frac{1}{N} \sum_{i \in S} \frac{1}{f_i} (\hat{Y}_{i1} - \hat{Y}_{i0}) \right) \]

and

\[ \hat{V}_1^t = \hat{V}_{svy} \left( \frac{1}{N_1} \sum_{i \in S} \frac{Z_i}{f_i} (\hat{Y}_{i1} - \hat{Y}_{i0}) \right) \]
where \( \hat{Y}_{it1} \) and \( \hat{Y}_{it0} \) are defined as above depending on the estimator and \( \hat{V}_{svy} \) is the survey-design specific estimated variance around the estimated treatment effect \( \hat{\Delta} \). We don’t know the true functions \( \mu_1(X_i) \) and \( \mu_0(X_i) \), but we estimate their difference using our observed matched sets. The design specific variances are easily obtainable from common statistical computing packages. This part of the variance treats the estimated treatment effects in each matched set as if they are the true observed effects rather than estimated. When calculating \( \hat{V}_1 \) or \( \hat{V}_t^t \) for full matching estimators, we use the same estimated treatment effect for all observations in the same matched set. This will deflate our estimate of \( \hat{V}_1 \) or \( \hat{V}_t^t \) because we are using an effective sample size of \( n \) or \( n_1 \) instead of the number of matched sets \( J \) from which we actually get estimates of the treatment effect. To compensate for this fact add clusters that represent the matched sets as part of the survey design when calculating \( \hat{V}_1 \) or \( \hat{V}_t^t \). This will give a more accurate estimate of the variance because it captures the fact that all estimates within a cluster are the same.

\( E_M \) is the weighted average of the residuals of the conditional treatment effect. The variance of \( E_M \) when \( \hat{\Delta}_M = \hat{\Delta}_{match, full, svy} \) is written as

\[
V(E_M) = V \left( \frac{1}{N} \sum_{i \in S} \frac{1}{f_i} \left( Z_i \left( \frac{W_i^A}{W_i^T} \right) - (1 - Z_i) \left( \frac{W_i^A}{W_i^C} \right) \right) \right),
\]

which we estimate by

\[
\hat{V}_2 = \frac{1}{N^2} \sum_{i \in S} \frac{1}{f_i^2} \left( Z_i \left( \frac{W_i^A}{W_i^T} \right)^2 - (1 - Z_i) \left( \frac{W_i^A}{W_i^C} \right)^2 \right) \hat{\sigma}^2(X_i, Z_i),
\]

where \( \hat{\sigma}^2(X_i, Z_i) \) is an estimate of the conditional variance of \( Y \). Abadie and Imbens [1] estimate \( \sigma^2(X_i, Z_i) \) separately for each observation \( i \) by matching similar observations within each treatment group. This can be computationally intensive, and we instead assume \( \sigma^2(X, Z) \) is constant across \( X \) and \( Z \). Taking advantage of
the matched sets with two or greater treated or two or greater control observations, we estimate \( \sigma^2 \) by

\[
\hat{\sigma}^2 = \frac{1}{m_1 + m_0} \left( \sum_{j=1}^{J} I \left( \sum_{k=1}^{n_j} Z_{kj} \geq 2 \right) \sum_{k=1}^{n_j} Z_{kj} (Y_{kj} - \hat{\mu}_{1j})^2 \right) + \sum_{j=1}^{J} I \left( \sum_{k=1}^{n_j} (1 - Z_{kj}) \geq 2 \right) \sum_{k=1}^{n_j} (1 - Z_{kj}) (Y_{kj} - \hat{\mu}_{0j})^2,
\]

where

\[
m_1 = \sum_{j=1}^{J} I \left( \sum_{k=1}^{n_j} Z_{kj} \geq 2 \right) \sum_{k=1}^{n_j} Z_{kj}, \text{ and}
\]

\[
m_0 = \sum_{j=1}^{J} I \left( \sum_{k=1}^{n_j} (1 - Z_{kj}) \geq 2 \right) \sum_{k=1}^{n_j} (1 - Z_{kj}).
\]

We note that for 1:1 matching there will be no matched sets with 2 or greater treated or control units, so this estimator will not be useful, but with our other matching procedures we anticipate that in most situations there will be a sufficient number of matched sets with 2 or greater treated or control units to get an appropriate estimate of \( \sigma^2 \). In the case of 1:1 matching we can either use the estimator suggested by Abadie and Imbens [1] or we can ignore \( \hat{V}_2 \).

Putting \( \hat{V}_1 \) and \( \hat{V}_2 \) together, we estimate the variance of \( \hat{\Delta}^t_{matchK.svy} \) by

\[
\hat{V}(\hat{\Delta}^t_{matchK.svy}) = \hat{V}_{svy} \left( \frac{1}{N_1} \sum_{i \in S} Z_i \left( Y_{i1} - \hat{Y}_{i0} \right) \right) + \frac{1}{N_1^2} \sum_{i \in S} \frac{1}{f_i^2} \left( Z_i - (1 - Z_i) \left( \frac{W_i^T}{W'_i} \right)^2 \right) \hat{\sigma}^2.
\]

(4.11)
The estimated variance of \( \hat{\Delta}_{\text{matchfull.svy}} \) is
\[
\hat{V}(\hat{\Delta}_{\text{matchfull.svy}}) = \hat{V}_{\text{svy}} \left( \frac{1}{N} \sum_{i \in S} \frac{1}{f_i} (\hat{Y}_{i1} - \hat{Y}_{i0}) \right) + \frac{1}{N^2} \sum_{i \in S} \frac{1}{f_i^2} \left( Z_i \left( \frac{W_i^T}{W_i} \right)^2 - (1 - Z_i) \left( \frac{W_i^T}{W_i} \right)^2 \right) \hat{\sigma}^2. \tag{4.12}
\]

Lastly, the estimated variance of \( \hat{\Delta}_{t\text{matchfull.svy}} \) is
\[
\hat{V}(\hat{\Delta}_{t\text{matchfull.svy}}) = \hat{V}_{\text{svy}} \left( \frac{1}{N_1} \sum_{i \in S} \frac{1}{f_i} (\hat{Y}_{i1} - \hat{Y}_{i0}) \right) + \frac{1}{N_1^2} \sum_{i \in S} \frac{1}{f_i^2} \left( Z_i - (1 - Z_i) \left( \frac{W_i^T}{W_i} \right)^2 \right) \hat{\sigma}^2. \tag{4.13}
\]

### 4.5 Evaluating Balance

As with weighting and stratification estimators an important step in a matching analysis is checking the balance between the treated and control covariate distributions after estimating the propensity score and creating matches. Recall equations (2.30) and (2.31) that are used to evaluate the balance for continuous and categorical covariates. In a matching-estimator context with survey data, the before-matching balance is calculated as it was for weighting and stratification estimators. We use the sampling weights to calculate the weighted versions of (2.30) and (2.31) as we described in Section 2.5. This should be calculated using all treated and control observations in the sample, not just ones in the matched sample.

The after-matching balance is calculated using not only the the survey weights for a particular observation, but also the information from the matched set from which an observation belongs. For the estimator \( \hat{\Delta}_{\text{matchfull.svy}} \) (4.4), a treated observation \( i \) contributes a weight of \( \frac{1}{f_i} \left( \frac{W_i^A}{W_i^T} \right) \) to the estimator, while a control observation \( i \) contributes weight \( \frac{1}{f_i} \left( \frac{W_i^A}{W_i^C} \right) \) to the estimator. \( W_i^T \) is the sum of the treated survey weight in the matched set from which observation \( i \) belongs, \( W_i^C \) is the sum of the control
survey weight in the matched set from which observation $i$ belongs, and $W_i^d$ is the sum of the treated and control survey weight in the matched set from which observation $i$ belongs. For the estimator $\hat{\Delta}^t_{\text{matchfull.svy}}$ (4.6) the contributed weights are $\frac{1}{f_i}$ for treated observations and $\frac{1}{f_i} \left( \frac{W_i^T}{W_i^C} \right)$ for control observations. Lastly, for the estimator $\hat{\Delta}^t_{\text{matchK.svy}}$ (4.2), the weights are also $\frac{1}{f_i}$ for treated observations and $\frac{1}{f_i} \left( \frac{W_i^T}{W_i^C} \right)$ for control observations which are included in the match. Evaluating the balance using the standardized differences with these weights gives us the after-matching balance.

4.6 Simulation Studies

We create simulations to study the properties of our proposed matching estimators and compare them to our previous weighting estimators. Two different simulation setups are used depending on the estimators of interest. The first simulation is designed to study estimators of $\Delta_{PATE}$. We use the same exact scenario as we did in Section 2.8 and compare our proposed full matching estimator $\hat{\Delta}_{\text{matchfull.svy}}$ with its naive version $\hat{\Delta}_{\text{matchfull}}$ and weighting estimator $\hat{\Delta}_{\text{ipw3.svy}}$. As before, this simulation uses two continuous covariates and a binary covariate which are associated with both the outcome and treatment selection. Four degrees of treatment-effect heterogeneity are used, as are two values of $\beta$, $\beta_{\text{mod}}$ and $\beta_{\text{str}}$, which represent the strength of association between the covariates and the treatment group. A sample size of 1,000 is drawn from a population of 100,000 using four sampling strata of different sizes. The propensity score model is correctly specified for all estimators. The naive full matching estimator uses matches based on the estimated propensity score from a logistic regression model, which does not use the survey weights, while $\hat{\Delta}_{\text{matchfull.svy}}$ uses matches based on the estimated propensity score from a weighted logistic regression model. While we did calculate $\hat{\Delta}_{\text{matchfull.svy}}$ using the un-weighted estimated propensity scores, we
do not show the results here as there is no discernible difference in the bias or the variance.

The second simulation is designed to investigate the properties of estimators of $\Delta_{PATT}$. We compare $\hat{\Delta}_{matchk.svy}$, its naive version $\hat{\Delta}_{matchk}$, $\hat{\Delta}_{matchfull.svy}$, and its naive estimator $\hat{\Delta}_{matchfull}$. Both 1:K matching estimators match two control observations to each treated observation. This simulation setup is similar to the ones before with two exceptions. First, we modify the function used to generate the true propensity score. Previously the true propensity score was given by

$$e(X, \beta) = (1 + \exp(-(0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3)))^{-1},$$

where $\beta_{str} = (0.6, -0.6, 0.6)'$ and $\beta_{mod} = (0.3, -0.3, 0.3)'$, but now the true propensity score is given by

$$e(X, \beta) = (1 + \exp(-(2 - 0.3X_1 - 0.1X_2 + 0.1X_3)))^{-1}.$$

This shifts the whole distribution of the propensity score towards zero and makes it much more likely that an observation is in the control group rather than the treated group, which also has the effect of increasing the ratio of control to treated observations. We make this change to the simulation so that it is more realistic to a situation where 1:K matching might work well. Recall the discussion in Section 4.1.1 concerning the relative size of the treated and control samples as well as the distribution of the propensity scores. About 13.5% of the population is treated in this situation compared with the approximately 53% of the population that is treated in the previous simulations. The second change we make to the setup concerns the heterogeneity of the treatment effect. We double the magnitude of the coefficient of the interaction $\zeta$, so that $\zeta_0 = 0, \zeta_{sm} = 0.5, \zeta_{mod} = 1$, and $\zeta_{str} = 2$. This modification creates a larger
difference between the sample average treatment effect in the treated and the population average treatment effect in the treated. Treated observations are more alike in terms of their covariates, compared with the entire population, so the previous interaction coefficients did not create very large differences between the population and sample average effects.

The estimated variance of each survey-adjusted estimator is calculated as proposed earlier in the chapter, while the estimated variance of the naive matching estimators are calculated by the variance of the matched set differences. In the case of the $\Delta_{PATE}$ simulation, the variance of $\hat{\Delta}_{ipw3.svy}$ is estimated using the formulas in Section 2.6.2. The estimated variances are used to create the 95% confidence intervals for each estimator assuming the estimator is normally distributed.

Figure 4.1 shows the percent bias of the full matching estimators compared with that of the weighting estimator $\hat{\Delta}_{ipw3.svy}$ for the $\Delta_{PATE}$ simulation. Similarly to naive weighting and stratification estimators in previous simulations, the bias of the naive estimator $\hat{\Delta}_{matchfull}$ increases as the heterogeneity of the treatment effect increases. Meanwhile, the bias of the survey-adjusted estimator $\hat{\Delta}_{matchfull.svy}$ stays approximately the same, near zero.

The mean-squared error of the estimators for the $\Delta_{PATE}$ simulation are shown in figure 4.2. Interestingly, our adjusted matching estimator $\hat{\Delta}_{matchfull.svy}$ does as well or better in terms of MSE than $\hat{\Delta}_{ipw3.svy}$, which did the best of all weighting and stratification estimators. The gains for $\hat{\Delta}_{matchfull.svy}$ are more apparent under the $\beta = \beta_{str}$ scenario, where the initial treatment and control groups are more different from each other than in the $\beta = \beta_{mod}$ scenario.

We see the coverage of the 95% confidence intervals created by each of the three estimators in Figure 4.3. The coverage of $\hat{\Delta}_{matchfull.svy}$ is below the nominal level
in the $\beta = \beta_{str}$ scenario, but still does about as well or better than $\hat{\Delta}_{ipu3.svy}$. In the $\beta = \beta_{mod}$ scenario, the coverage of $\hat{\Delta}_{matchfull.svy}$ is slightly larger than in the other scenario, but still below the nominal level. As expected because of its bias, the coverage of the naive full matching estimator is very poor when the heterogeneity of the treatment effect is large.

The $\Delta_{PATE}$ simulation shows that our proposed full matching estimator not only adjusts for the survey design in terms of the bias, but it also does as well or better than any weighting or stratification estimator in terms of efficiency. When the propensity score is properly specified and we can find good matches for the sampled observations there may be some advantage to the extra effort of matching the observations.

The distribution of the propensity score in the $\Delta_{PATT}$ simulation is considerably different than in previous simulations because 1:K matching estimators need a large ratio of control observations to treated observations to work well. Figure 4.4 shows a histogram of the propensity score distributions of the sample from one iteration of the simulation. Note that while the propensity scores are mostly less than 0.3, there are a small number of observations with larger estimated propensity scores. As we discussed earlier, 1:K matching methods have a difficult time finding good matches for all treated observations with larger propensity scores. Here, since we are creating 1:2 matched sets, treated observations with a propensity score greater than one third may not be able to find enough suitable controls for matching.

Figure 4.5 shows the percent bias and MSE of the estimators. When the treatment effect is homogeneous, all the estimators are approximately unbiased with the exception of $\hat{\Delta}^t_{matchK}$, which is slightly biased. This is likely due to the treated observations with large propensity scores for which we could not find good matches. As the treatment heterogeneity increases, so does the bias of the naive estimators.
Meanwhile, the survey-adjusted estimators remain approximately unbiased. The bias and MSE for $\hat{\Delta}^t_{\text{matchK,svy}}$ and $\hat{\Delta}^t_{\text{matchfull,svy}}$ are approximately the same, with neither having a distinct advantage.

The coverage of the estimators is shown in Figure 4.6. Both survey-adjusted matching estimators $\hat{\Delta}^t_{\text{matchfull,svy}}$ and $\hat{\Delta}^t_{\text{matchK,svy}}$ have coverage above the nominal level, with $\hat{\Delta}^t_{\text{matchfull,svy}}$ being more conservative. In the $\Delta_{PATE}$ simulation, the coverage of $\hat{\Delta}_{\text{matchfull,svy}}$ was below the nominal level (Figure 4.3), while here the coverage of $\hat{\Delta}_{\text{matchfull,svy}}$ is above its nominal level. This may be in part due to how our proposed variance estimators compute $\hat{V}_{svy} \left( \frac{1}{N_1} \sum_{i \in S_1} f_i \left( \hat{Y}_{i1} - \hat{Y}_{i0} \right) \right)$ and $\hat{V}_{svy} \left( \frac{1}{N} \sum_{i \in S} f_i \left( \hat{Y}_{i1} - \hat{Y}_{i0} \right) \right)$, respectively in equations (4.12) and (4.13). The first is computed across all units in the sample and, while the second is computed across only the treated units and thus is larger.

This simulation shows that our proposed survey-adjusted methods correct for the bias in the naive methods when the treatment effect is heterogeneous, as in past simulations. The coverage of both proposed estimators is larger than the nominal level, which indicates that we are overestimating the variance. Refinements to the variance estimators should be explored, but the estimators are at least conservative. Neither $\hat{\Delta}^t_{\text{matchfull,svy}}$ nor $\hat{\Delta}^t_{\text{matchK,svy}}$ performed better than the other in this simulation, but recall that $\hat{\Delta}^t_{\text{matchK,svy}}$ may become biased when treated observations have large propensity scores because we may not be able to find good matches for all the observations.
4.7 Tables and Figures

Figure 4.1: Percent bias of propensity score weighting and matching estimators for $\Delta_{PATE}$.
Figure 4.2: MSE of propensity score weighting and matching estimators for $\Delta_{PATE}$.

Figure 4.3: Coverage of 95% confidence intervals created by propensity score weighting and matching estimators for $\Delta_{PATE}$. 
Figure 4.4: Histogram of the estimated propensity scores in a sample from one iteration of the $\Delta_{PATT}$ simulation.

Figure 4.5: Percent bias and MSE of propensity score matching estimators for $\Delta_{PATT}$. 
Figure 4.6: Coverage of 95% confidence intervals created by propensity score matching estimators for $\Delta_{PATT}$. 
Chapter 5: Example: Health Reform Impact on Individual Health Outcomes

In this chapter we consider a real data application of our proposed methods. Using the 2012 Ohio Medicaid Assessment Survey (OMAS), a dual-frame telephone health survey representative of Ohio’s non-institutionalized adults, we attempt to estimate the effect of health insurance coverage on the self-rated health of adults 19-64 years of age whose 2012 income was between 138% and 400% of the Federal Poverty Level (FPL).

5.1 Introduction

As part of the 2010 Patient Protection and Affordable Care Act (ACA), state-based healthcare insurance exchanges were created to increase adults’ access to insurance coverage. While a majority of adults ages 19 through 64 years in Ohio have employer-sponsored health insurance (54.4% in 2012), approximately one out of every six adults ages 19 through 64 years was uninsured (17.3%) [44]. The healthcare insurance exchange allows adults who are uninsured or underinsured to buy affordable healthcare insurance. Starting in January 2014 persons with income up to 400% of the FPL qualified for tax credits to make purchasing insurance coverage more affordable [31]. In Ohio, starting January 1, 2014 Medicaid access was extended to adults ages 19 to 64 years between 0% and 138% of the FPL [32]. These two points
in conjunction create the “exchange population”, which is adults 19-64 years of age with incomes between 138% and 400% of the FPL.

Using the 2012 OMAS, we attempt to estimate the causal effect of having health insurance coverage by comparing adults who are chronically uninsured (those uninsured for at least the past year) with adults who have job-based or directly-purchased health insurance. We consider the outcome of self-rated health, both as a continuous variable and as a binary variable. Self-rated health is a strong predictor of mortality risk and is even independently predictive given physician assessments of health [20].

5.2 Data

The 2012 OMAS is a population-based survey that measured the health insurance coverage, health status, and health care experiences of Ohio’s adult and child populations. The design of the survey is a stratified simple random sample of telephone numbers in Ohio, which included both landline and cell-phone numbers. The design included 105 unique strata which included one stratum for each non-metropolitan county (81 strata), three strata for each of the seven metropolitan counties divided into high-, medium-, and low-density areas where African Americans live based on Census tract (21 strata), an Asian and a Hispanic surname stratum (2 strata), and a single statewide cell phone stratum (1 stratum). The survey took place from May to October 2012 during which 22,929 Ohio residents were interviewed.

5.3 Methods

We create treatment and control groups consisting of 3,439 and 483 adults respectively with reported or imputed incomes between 138% and 400% FPL and reported or imputed ages between 19 and 64 years. Here treatment indicates having insurance,
while control indicates being uninsured. The treated group only includes those with job-based or directly-purchased insurance, not Medicaid or Medicare, and further excluded those who are part of the military. Persons with Medicaid, Medicare, or who are part of the military will have insurance at this point in time, and thus both their potential outcomes do not exist. The control group included only those who indicated they have been uninsured for at least the past year. A small number, 13 control and 91 treated, of observations are removed because of missing covariates.

After discussions with content experts, we included 16 covariates available to us in the survey in the analysis as confounders. These variables are: race as a five category variable, age as a continuous variable, sex, education as a three category variable, income as a percent of the federal poverty level as a four category variable, county type as a four category variable (Appalachian, Metropolitan, Rural Non-Appalachian, Suburban), a variable that approximates disability, working status as a three category variable, children in the house (0, 1, 2, 3 or more), marital status (Married, Single or Separated, Unmarried Couple), smoking status (Yes or No), drinking status (Yes or No), mental health distress (Yes or No), ever been told has cancer (Yes or No), ever been told has diabetes (Yes or No), and ever been told has high blood pressure (Yes or No).

All possible confounders above were included in the propensity score model except for ever been told has cancer, ever been told has diabetes, and ever been told has high blood pressure, because these variables were sufficiently balanced in the initial comparisons of treated and control groups and after the propensity score adjustment. We built the propensity score model using weighted logistic regression and included several two-way interactions to achieve better balance between groups. We compared our estimators to those using propensity scores created from non-weighted logistic
regression under the same model. The logistic regression model was created with
the weighting estimators $\hat{\Delta}_{ipw2,svy}$ and $\hat{\Delta}_{ipw3,svy}$ in mind, meaning that we adjusted
the model in order to optimize the balance after weighting for these estimators. We
chose to focus on weighting estimators, rather than stratification, hybrid, or matching
estimators because of the nice theoretical properties we showed in Chapter 2, but we
do present results for all four types of estimators. If we attempted to optimize the
balance for other estimators, we likely would make different choices for terms in the
model, but we chose to create a single estimated propensity score for all estimators to
try to avoid confusion. Because of this fact, the balance metrics for some matching
and stratification estimators are less than ideal.

We first consider self-rated health status on a five-point scale as a continuous
variable, which is imputed for missing values. We then treat self-rated health as bi-
nary (fair/poor vs. excellent/very good/good). We estimate the treatment effect and
variance estimates using naive and proposed weighting, stratification, and matching
estimators. We use weighting estimators $\hat{\Delta}_{ipw2}, \hat{\Delta}_{ipw3}, \hat{\Delta}_{ipw2,svy}$, and $\hat{\Delta}_{ipw3,svy}$
which includes using both weighted and unweighted propensity score estimates for
$\hat{\Delta}_{ipw2,svy}$. For all stratification estimators we use five propensity score strata, and we
include estimators $\hat{\Delta}_s, \hat{\Delta}_{h.ipw2}, \hat{\Delta}_{s.svy}$, and $\hat{\Delta}_{h.ipw2,svy}$. We compute matching estima-
tors $\hat{\Delta}^{t}_{matchK}, \hat{\Delta}^{t}_{matchfull}, \hat{\Delta}^{t}_{matchfull.svy}, \hat{\Delta}^{t}_{matchK.svy}, \hat{\Delta}^{t}_{matchfull.svy}$, and $\hat{\Delta}^{t}_{matchfull.svy}$. The
estimators $\hat{\Delta}^{t}_{matchK}$ and $\hat{\Delta}^{t}_{matchK.svy}$ matched three treated (insured) units to each
control unit and utilize a caliper of 0.25 standard errors of the estimated propensity score. These estimators estimate the average treatment effect in the control
(uninsured) population rather than the treated population. Matches were created
in R [43] using the pairmatch function in the package optmatch [11], which creates
matches that minimize the sum of the absolute differences in treated and control
matched estimated propensity scores. Matches for the full matching estimators utilized the \textit{fullmatch} function in the package \textit{optmatch} under the constraint that the matched sets should have at most a 25:1 ratio of treated to control units or control to treated units.

The weighting estimators and stratification estimators as well as the full matching estimators \( \hat{\Delta}_{matchfull, svy} \) and \( \hat{\Delta}_{matchfull} \) estimate the average effect of health insurance on self-rated health in the entire population. Meanwhile, the matching estimators \( \hat{\Delta}_{matchK, svy} \), \( \hat{\Delta}_{matchK, svy}^t \), \( \hat{\Delta}_{matchfull}^t \), and \( \hat{\Delta}_{matchfull, svy}^t \) estimate the average effect in the uninsured population. When self-rated health is treated as binary, the point estimates are interpreted as the estimated difference in prevalence of fair/poor self-rated health if all persons in the population, or all persons in the control population, had health insurance compared to all persons being uninsured. The interpretation when self-rated health is viewed as continuous is the estimated difference in average rating of fair/poor self-rated health in points if all persons in the population, or all persons in the control population, had health insurance compared to all persons being uninsured.

\section*{5.4 Results}

Figure 5.1 shows the smoothed distribution of the estimated propensity score for the treated and control groups, as well as a box plot of the propensity scores. In this figure the propensity score is estimated with weighted logistic regression using the survey weights. We see that the majority of observations have a propensity score larger than 0.5 and there are only limited observations on the lower end of the propensity score distribution. Here we see the distributions of the propensity scores within each treatment group, rather than in total. Recall that there are more than seven times as many treated observations as there are control observations.
We plot the estimated propensity scores from the same model using unweighted and weighted logistic regression in Figure 5.2. This figure is very different from Figure 2.3, the same comparison using simulated data. Using real data, we do not know the true propensity score model, and the survey weights here are more variable than the ones in our simulation. As a result, we expect the treatment effect estimates and the implied balance of estimators using weighted and unweighted propensity scores to be less similar than in the earlier simulation.

Figures 5.3 and 5.4 show the balance of confounders between treated and control groups before and after propensity score weighting using weights implied by $\hat{\Delta}_{ipw2.svy}$. In Figure 5.3 we use the estimated propensity scores from the weighted logistic regression model, and in Figure 5.4 we use the estimated propensity scores from the unweighted logistic regression model. The initial distributions of the confounders are quite different for the treatment groups with education, income, marital status, and smoking status being the most different. The absolute standardized differences of the confounders after weighting all fall below the 10% rule-of-thumb threshold in Figure 5.3, which uses the weighted propensity scores. While the balance using the weighted propensity scores is clearly superior to the balance when using the unweighted propensity scores in these figures, it should be noted that the common propensity score model for all estimators is created explicitly to optimize the balance of $\hat{\Delta}_{ipw2.svy}$ and $\hat{\Delta}_{ipw3.svy}$ using the weighted logistic regression model. It is possible that there exists a different propensity score model which leads to better balance using the unweighted propensity scores. The balance for the estimator $\hat{\Delta}_{ipw3.svy}$ is almost exactly the same as that of $\hat{\Delta}_{ipw2.svy}$, so we do not include the plot.

In Figures 5.5 and 5.6 we see the balance implied by the stratification estimator $\hat{\Delta}_{s.svy}$ and the hybrid estimator $\hat{\Delta}_{h.ipw2.svy}$. For both of these estimators it is necessary
to check the balance in each propensity score stratum. Neither does well in balancing the confounders in all the strata. In particular Stratum 5 in both figures appears to be problematic. Stratum 5 contains the observations with the largest propensity scores, and the ratio of treated to control observations is very large. In fact, this stratum contains only 7 control observations but 778 treated observations. Reducing the number of strata to four does not alleviate this problem. The data set does not appear to be a good candidate for stratification estimators in general. Using a larger number of strata should lead to better balance, but in this case we found that it leads to too few observations in each of the propensity score stratum making balance of the many covariates difficult.

The balance implied by the matching estimators $\hat{\Delta}_{\text{matchK.svy}}^t$ and $\hat{\Delta}_{\text{matchfull.svy}}^t$ are shown in Figures 5.7 and 5.8. In both cases, the after-matching balance is less than the 10% absolute standardized difference rule of thumb for most, but not all, of the variables. This is in part because our propensity score model is not optimized for these estimators.

Self-rated health status is measured on a five-point scale with 1 representing excellent health and 5 representing poor health. The average self-rated health status in our insured population is 2.311, while in our uninsured population it is 2.567. The estimated prevalence of poor-fair self-rated health status is 0.114 in our insured population and 0.200 in the uninsured population. The estimated effect of health insurance on self-rated health as a continuous variable for each type of estimator, its variance, and its 95% confidence interval is presented in Table 5.1. Table 5.2 shows the same estimates when we treat self-rated health as a binary variable (fair or poor self-rated health =1). When the outcome is binary, these estimates can be interpreted as estimated differences in prevalence.
For the most part, point estimates for both outcomes of interest are negative, indicating that health insurance is beneficial. The initial difference in means without adjustment when the outcome is continuous (Table 5.1) is much larger than after propensity score adjustment. This indicates that much of the initial difference in means can be explained by differences in the distributions of covariates between treated and control groups. Naive weighting and stratification estimates are generally smaller in magnitude than their survey-adjusted counterparts, while the opposite is true for matched estimates. Comparing the results from different types of estimators is complicated by the fact that they represent different degrees of balance between treatment groups. We can’t say for sure whether the differences in estimates is a result of the estimators themselves or the different degrees of balance that the estimators represent.

We focused on weighting estimators when building the propensity score model, and we believe \( \hat{\Delta}_{ipw, svy} \) represents our best estimate of the causal effect of insurance status because it achieves the best balance. We estimated that having health insurance coverage decreases the self-rated health score by 0.1252 (95% CI [0.0059, 0.2446]) points on average for our population. This is the estimated average difference if everyone in the population received health insurance coverage compared to if no one in the population received health insurance coverage. We estimate that having health insurance coverage decreases the prevalence of fair-poor health on average in our population by 0.0723 (95% CI [0.0253, 0.1193]).

The estimated population average treatment effect in the control population (uninsured) from the matching estimators yields interesting results. It appears that the estimated average effect in the uninsured population is smaller in magnitude than in overall population and thus the insured population.
5.5 Conclusions

As expected, we estimate a beneficial effect of health insurance coverage on self-rated health on average for adults 19-64 years of age with incomes between 138% and 400% of the federal poverty level. The added stress of not having insurance health insurance might decrease one’s health directly, and in addition health insurance provides greater access to medical care which in turn may contribute to one’s health through preventative or reactive encounters with the health system.

While the interpretation of self-rated health as a continuous outcome is difficult, when used as a binary variable it is easier to understand the impact. We estimate that having health insurance coverage decreases the prevalence of fair-poor health on average in our population by 0.0723 (95% CI [0.0253, 0.1193]). To further put this difference in perspective, the estimated prevalence if no persons received health insurance compared to all persons would be 0.1928 to 0.1205, respectively. In the context of the size of our population, this represents a difference of approximately 147,581 people. The difference between 0.1928 to 0.1205 represents an estimated 37.5% change in the prevalence of fair-poor health.

Our study has several limitations. First of all, we are assuming no unmeasured confounding in our data, which is unlikely to be true. Another limitation of our analysis is the temporal aspect of health insurance coverage. Having health insurance is not a static event at one point in time, or over a specific length of time. Over the course of one’s lifetime, one may have and not have insurance coverage during certain periods of time. In contrast, our analysis treats insurance status as a static treatment, which is of course not true. Next, the question of pre-existing conditions merits discussion. Prior to 2014, it was possible for an insurance company to deny coverage to adults due to a pre-existing condition such as heart disease or cancer.
This reverses the direction of the causality between insurance status and self-rated health because now self-rated health determines insurance status. Unfortunately, we are unable to detect this using OMAS data. However, having a pre-existing condition did not always preclude one from getting health insurance. We did include indicators of disability and mental health distress in our propensity score model and monitored the balance of reported cancer, diabetes, and high blood pressure indicators between treatment and control groups. These variables should capture at least some of the increased probability of being uninsured because of a pre-existing condition. A final limitation of this analysis is the fact that not all health insurance coverage plans are the same. A person who does have health insurance coverage may still have poor access to medical care because of limited coverage options. There is no way of determining whether some insurance coverage is better or worse than others using this data.

While this analysis attempted to estimate the effect of health insurance on average in the entire population, additional work is needed to determine which groups benefit most from insurance coverage. Other outcomes may also be of interest with regard to the effect of health insurance including emergency department usage and expenditures. In addition to analysis concerning subpopulations and other outcomes of interest, better methodology in regards to missing data should be incorporated into this analysis. We used several variables for which missing values had been previously imputed for the survey and we discarded several observations because of missing covariates. In future work it would be best to incorporate a multiple imputation approach to account for added uncertainty due to missing data, which is often a concern with survey data.
5.6 Tables and Figures

Figure 5.1: The smoothed distributions and box and whiskers plot of the estimated propensity score for treated and control groups.
Figure 5.2: Estimated propensity scores using unweighted and weighted logistic regression.

Figure 5.3: The balance of treated and control groups before and after propensity score weighting using weights implied by \( \Delta_{ipu2.,svy} \), using propensity scores from weighted logistic regression.
Figure 5.4: The balance of treated and control groups before and after propensity score weighting using weights implied by $\hat{\Delta}_{ipw2,svy}$, using propensity scores from a unweighted logistic regression.

Figure 5.5: The balance of treated and control groups before and after stratification using $\hat{\Delta}_{s,svy}$.
Figure 5.6: The balance of treated and control groups before and after stratification and weighting using $\hat{\Delta}_{h,ipw2.svy}$.

Figure 5.7: The balance of treated and control groups before and after matching using $\hat{\Delta}_{matchK.svy}$.
Figure 5.8: The balance of treated and control groups before and after matching using \( \hat{\Delta}_{\text{match full .svy}} \).

<table>
<thead>
<tr>
<th>Race/Ethn–White</th>
<th>Race/Ethn–Black</th>
<th>Race/Ethn–Asian</th>
<th>Race/Ethn–Other</th>
<th>Age (Continuous)</th>
<th>Gender</th>
<th>Not Working</th>
<th>Work &lt;35 hrs/wk</th>
<th>Work 35+ hrs/wk</th>
<th>High School or Below</th>
<th>Some College or Associate's</th>
<th>College Grad or above</th>
<th>Disability</th>
<th>Income 139% – 200% FPL</th>
<th>Income 201% – 250% FPL</th>
<th>Income 251% – 300% FPL</th>
<th>Income 301% – 400% FPL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Before Matching</td>
<td>After Matching</td>
<td>Before Matching</td>
<td>After Matching</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 5.1: Point estimates, estimated variance, and estimated 95% confidence intervals for the effect of health insurance on self-rated health as a continuous variable.

<table>
<thead>
<tr>
<th>Estimator</th>
<th>Point Estimate</th>
<th>Est.Var</th>
<th>95% CI LB</th>
<th>95% CI UB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Initial Difference in Means</td>
<td>-0.2559</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Naive IPW2</td>
<td>-0.0580</td>
<td>0.0060</td>
<td>-0.2094</td>
<td>0.0935</td>
</tr>
<tr>
<td>Naive IPW3</td>
<td>-0.0522</td>
<td>0.0059</td>
<td>-0.2026</td>
<td>0.0983</td>
</tr>
<tr>
<td>Survey IPW2*</td>
<td>-0.0952</td>
<td>0.0102</td>
<td>-0.2935</td>
<td>0.1031</td>
</tr>
<tr>
<td>Survey IPW2</td>
<td>-0.1252</td>
<td>0.0037</td>
<td>-0.2446</td>
<td>-0.0059</td>
</tr>
<tr>
<td>Survey IPW3</td>
<td>-0.1294</td>
<td>0.0038</td>
<td>-0.2504</td>
<td>-0.0084</td>
</tr>
<tr>
<td>Naive Stratified</td>
<td>-0.0701</td>
<td>0.0089</td>
<td>-0.2551</td>
<td>0.1149</td>
</tr>
<tr>
<td>Naive Hybrid IPW2</td>
<td>-0.0594</td>
<td>0.0058</td>
<td>-0.2088</td>
<td>0.0901</td>
</tr>
<tr>
<td>Survey Stratified</td>
<td>-0.0899</td>
<td>0.0110</td>
<td>-0.2951</td>
<td>0.1152</td>
</tr>
<tr>
<td>Survey Hybrid IPW2</td>
<td>-0.1187</td>
<td>0.0043</td>
<td>-0.2466</td>
<td>0.0092</td>
</tr>
<tr>
<td>Naive Full Match</td>
<td>-0.0904</td>
<td>0.0031</td>
<td>-0.1998</td>
<td>0.0191</td>
</tr>
<tr>
<td>Survey Full Match</td>
<td>-0.0866</td>
<td>0.0048</td>
<td>-0.2227</td>
<td>0.0495</td>
</tr>
<tr>
<td>Naive 1:3 Match (PATC)</td>
<td>-0.1035</td>
<td>0.0031</td>
<td>-0.2131</td>
<td>0.0061</td>
</tr>
<tr>
<td>Survey 1:3 Match (PATC)</td>
<td>-0.0112</td>
<td>0.0056</td>
<td>-0.1576</td>
<td>0.1353</td>
</tr>
<tr>
<td>Naive Full Match (PATC)</td>
<td>-0.1385</td>
<td>0.0039</td>
<td>-0.2616</td>
<td>-0.0154</td>
</tr>
<tr>
<td>Survey Full Match (PATC)</td>
<td>0.0526</td>
<td>0.0092</td>
<td>-0.1353</td>
<td>0.2404</td>
</tr>
</tbody>
</table>

*Using propensity scores from the logistic regression model without survey weights
Table 5.2: Point estimates, estimated variance, and estimated 95% confidence intervals for the effect of health insurance on the prevalence of fair/poor self-rated health.

<table>
<thead>
<tr>
<th>Estimator</th>
<th>Point Estimate</th>
<th>Est.Var</th>
<th>95% CI LB</th>
<th>95% CI UB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Initial Difference in Means</td>
<td>-0.0858</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Naive IPW2</td>
<td>-0.0388</td>
<td>0.0007</td>
<td>-0.0912</td>
<td>0.0136</td>
</tr>
<tr>
<td>Naive IPW3</td>
<td>-0.0384</td>
<td>0.0007</td>
<td>-0.0908</td>
<td>0.0141</td>
</tr>
<tr>
<td>Survey IPW2*</td>
<td>-0.0645</td>
<td>0.0009</td>
<td>-0.1232</td>
<td>-0.0057</td>
</tr>
<tr>
<td>Survey IPW2</td>
<td>-0.0723</td>
<td>0.0006</td>
<td>-0.1193</td>
<td>-0.0253</td>
</tr>
<tr>
<td>Survey IPW3</td>
<td>-0.0748</td>
<td>0.0006</td>
<td>-0.1221</td>
<td>-0.0274</td>
</tr>
<tr>
<td>Naive Stratified</td>
<td>-0.0362</td>
<td>0.0010</td>
<td>-0.0987</td>
<td>0.0262</td>
</tr>
<tr>
<td>Naive Hybrid IPW2</td>
<td>-0.0416</td>
<td>0.0007</td>
<td>-0.0940</td>
<td>0.0108</td>
</tr>
<tr>
<td>Survey Stratified</td>
<td>-0.0490</td>
<td>0.0011</td>
<td>-0.1147</td>
<td>0.0167</td>
</tr>
<tr>
<td>Survey Hybrid IPW2</td>
<td>-0.0716</td>
<td>0.0006</td>
<td>-0.1206</td>
<td>-0.0226</td>
</tr>
<tr>
<td>Naive Full Match</td>
<td>-0.0361</td>
<td>0.0004</td>
<td>-0.0749</td>
<td>0.0027</td>
</tr>
<tr>
<td>Survey Full Match</td>
<td>-0.0259</td>
<td>0.0005</td>
<td>-0.0683</td>
<td>0.0165</td>
</tr>
<tr>
<td>Naive 1:3 Match (PATC)</td>
<td>-0.0366</td>
<td>0.0005</td>
<td>-0.0788</td>
<td>0.0056</td>
</tr>
<tr>
<td>Survey 1:3 Match (PATC)</td>
<td>-0.0187</td>
<td>0.0007</td>
<td>-0.0716</td>
<td>0.0342</td>
</tr>
<tr>
<td>Naive Full Match (PATC)</td>
<td>-0.0508</td>
<td>0.0006</td>
<td>-0.0977</td>
<td>-0.0039</td>
</tr>
<tr>
<td>Survey Full Match (PATC)</td>
<td>-0.0167</td>
<td>0.0012</td>
<td>-0.0832</td>
<td>0.0497</td>
</tr>
</tbody>
</table>

*Using propensity scores from the logistic regression model without survey weights
Chapter 6: Discussion and Future Work

In this dissertation we developed a framework and estimators for estimating causal treatment effects from observational complex survey data using propensity score adjustments. We found that naive propensity score estimators, which do not take into account the survey design, will be biased when the treatment effect is heterogeneous. We proposed weighting, stratification, matching, and a new class of hybrid estimators for population average treatment effects that take into account the survey design along with estimators of their respective variances. In general, the topic of propensity score estimators to estimate causal effects using complex survey data is not well researched, and existing methods are mostly ad-hoc without rigorous statistical justification. In addition, no research has been completed on how to properly estimate variances of such estimators.

Chapter 2 of this dissertation proposed a framework from which we thought about observing potential outcomes in a complex survey setting. This motivated our estimators throughout the dissertation. Later in the chapter we proposed and studied the properties of both inverse probability weighting estimators and stratification estimators. In Chapter 3, we proposed a new class of estimators which combined stratification and weighting techniques in an attempt to create a robust, unbiased estimator with small variance. We first developed the idea in the typical simple random sample case, and then translated those ideas to the complex survey setting. Chapter
proposed and discussed matching estimators that are analogous to 1:K and full matching that take into account the complex survey structure. Lastly, in Chapter 5, we used real survey data to demonstrate the proposed estimators in an example concerning the effect of health insurance coverage on self-reported health status.

6.1 Discussion

If a treatment effect is heterogeneous across different groups in a population and those groups are sampled at different rates in a survey, then any method that does not account for the sampling design or explicitly model the heterogeneity in the treatment effect will be biased. In this dissertation we focused on estimators with non-parametric outcome models, which means that accounting for the sampling design is essential. The framework which we proposed consists of two distinct sampling steps which ultimately lead to our sample data. The first step is the treatment selection and the second step is selection into the survey sample.

Like propensity score stratification estimators in the non-survey context, we proposed stratification estimators that stratify the observed data into $K$ groups based on their estimated propensity scores. Within each propensity score stratum, the goal is to have observed confounders independent of treatment assignment. If the observed covariates can be balanced between treated and control groups within each stratum, it allows us to estimate the treatment effect free of bias. We use survey weights in the estimation of the treatment effect within each stratum, and we also use them to estimate the proportion of the population that falls into each stratum. We showed that the stratification estimator is unbiased for the population average treatment effect under very strict assumptions.
The proposed weighting estimators use the joint probability of belonging to the treatment group and being sampled, to create weights for the treated and control observations that implied a weighted pseudo-population in which the measured confounders are balanced between groups. The weights allow us to estimate the treatment effect. The weighting estimators we proposed are consistent estimators of the population average treatment effect and in one case unbiased, under certain assumptions.

While similar propensity score estimators to these stratified and weighted estimators have been discussed previously in the literature, we contributed more theoretical results and extensive simulation on the topic. Additionally we addressed estimating the variances of stratified and weighting estimators using estimators that accommodate the complex survey design. Simulation results showed that as the treatment effect becomes more heterogeneous, naive stratification and weighting estimators of the population average treatment effect become increasingly biased, whereas our proposed estimators remained approximately unbiased or slightly biased for weighting and stratified estimators respectively. The variances of weighting estimators were smaller when using weighted logistic regression using the survey weights to estimate the propensity score rather than un-weighted logistic regression. This was true even when the propensity score model was properly specified.

Weighting estimators have attractive properties when the propensity score model is correctly specified, but when this is not the case they can be much more sensitive to model misspecification than stratified estimators. The hybrid estimators we proposed in Chapter 3 take on the property from inverse probability weighting of being consistent and nearly unbiased when the propensity score model is correct, but take on the robustness properties of stratification estimators when the propensity score model is incorrect. We developed these hybrid estimators first in the non-complex
survey setting and showed that they consistently estimate the average treatment effect when the propensity score model was correct. We then proposed corresponding hybrid estimators for the complex-survey context. In both cases simulations showed that hybrid estimators reached or improved on the efficiency of the best weighting estimator $\hat{\Delta}_{ipw3}$ or $\hat{\Delta}_{ipw3.svy}$.

We lastly proposed matching estimators and estimates of their variance that adjust for the survey design that are analogous to 1:K and full matching estimators. For these estimators, after matching on the propensity score, the survey weights of observations in each matched set are used both to estimate the treatment effect via a weighted average and to estimate the empirical distribution of the propensity score. When matching was exact, the probability of sampling was properly specified, and the treated and control observations shared a common support, we showed that the proposed 1:K and full matching estimators are unbiased. Full Matching was shown to be more flexible than 1:K matching because of its ability to match a variable number of control observations to a variable number of treated observations, and its ability to estimate the population average effect or the average effect in the treated and control populations. Our simulations showed that our proposed estimators $\hat{\Delta}_{match.full.svy}$ and $\hat{\Delta}_{matchK.svy}$ are approximately unbiased when the propensity score is correctly specified, whereas naive matching estimators became biased with increased treatment effect heterogeneity.

When trying to estimate the causal effect from some treatment using complex survey data the best propensity score estimator to use depends on the situation. The best place to start is with weighting estimators. We find that weighting estimators are the simplest to implement and when the propensity score is correctly specified, these estimators will be unbiased or nearly unbiased. However, weighting estimators
are more affected by misspecification of the propensity score and extreme weights than are other methods. Stratification methods will likely not remove all bias from an estimate, but they are more robust to misspecification of the propensity score. If this is a concern, we recommend the hybrid estimator $\hat{\Delta}_{h.ipw2}$.

Matching estimators are commonly used because of their ease of interpretation to non-statisticians, however they can be more difficult to implement than other estimators because of the matching step. We found that in some circumstances it was more difficult to create sufficient balance between the treated and control groups using matching estimators, and common statistical software packages that create matches were limited in the size of the groups they can match. Still, we found that full matching estimators performed as well or better than weighting estimators in our simulation. We do prefer using our proposed survey-adjusted full matching estimator compared to the proposed 1:K estimator because it is more flexible to a number of circumstances.

### 6.2 Limitations

In general, propensity score methods for estimating causal effects are limited by their ability to measure all relevant confounders and estimate the propensity score so that treatment selection is independent of the potential outcomes conditional on the propensity score. Our proposed methods are no different. Great care should be taken to identify the correct confounders for inclusion in the propensity score model, and the implied balance of the method should always be checked. Additionally, with our methods in a complex survey environment, we rely on the correct specification of the sampling weights. Propensity score methods are also limited in their ability
to produce a quality estimate if there is a lack of common support between the
distributions of the propensity score in the treated and control groups.

All of our proposed methods used a non-parametric outcome model, and while this
was a deliberate decision on our part, it can also lead to a lack of efficiency. Estimators
using an outcome model, such as a regression model, will be more efficient than any
of our proposed methods when that outcome model is correct. However, one of the
main advantages and reasons for using propensity score estimators is that we don’t
have to model the outcome correctly, which can be difficult to do when faced with a
situation as in Chapter 5 with a large number of possible confounders.

Another limitation is the lack of simulation study when the propensity score is
misspecified. Outside of Chapter 3, all our simulations used the correct model for the
propensity score. As we observed in the simulations in that chapter, stratification
estimators will likely be more robust to misspecification of the propensity score than
weighting estimators because they do not rely on the propensity score as directly. We
expect that matching estimators act more similarly to stratification estimators than
weighting estimators and will also be somewhat robust. We plan to conduct more
simulation studies in future work.

Survey data add an extra level of difficulty when attempting to balance the distri-
butions of treated and control observations after weighting, stratifying, or matching.
When sampling weights vary greatly, certain observations can be especially influential
in both the balance metrics and the estimate of the outcome. We found this to be
ture in the example in Chapter 5. Without including survey weights, it was much
easier to achieve the desired degree of balance.

In this work, we assumed the survey design did not include clustering as a major
part of the sampling design, as in the Ohio Medicaid Assessment Survey in Chapter 5.
While there are many surveys that fall into this category, it does not cover all possibilities. Designs that include clustering in their first stage will likely require additional modifications to our proposed methods in order to accommodate the clustering. We discuss possible solutions to this problem in the following section. Replicate weights are another type of survey design feature that we have not discussed in this dissertation. When the survey design characteristics are unknown to the researcher because of confidentiality reasons, replicate weights are often included to allow for variance estimation. We expect that estimating the variance components when the survey structure is needed using the replicate weights instead of using the design should also work well. The one exception might be the full matching estimators, for which we treated the matched sets as clusters for the estimates of variance.

Missing data are often a problem with survey data, and we have not addressed it at all up to this point. In order to maintain representativeness of the population average treatment effect it is best not to discard observations because of missing data. We discuss this issue further in the next section.

6.3 Future Work

Research on propensity score methods with complex survey data is relatively new, which leaves many potential areas for future work on this topic. The topics of cluster designs, missing data, and identifying heterogeneity are all possible extensions to our research.

6.3.1 Cluster Designs

When the sample design includes clustering at the first stage of sampling, the methods we proposed in this dissertation likely should not be used because the clustering creates dependence between observations for which are methods do not account.
Li et al. describe propensity score methods for estimating the average treatment effect from hierarchical data [24], which includes clusters. In their method, they estimate the propensity score using a cluster-level random effect model. Treatment effects are estimated within each cluster and then averaged with respect to the size of the cluster. Their work focuses on weighting estimators [24], but the methods should be able to be applied to stratification as well as matching estimators. For complex samples in which the first sampling stage is clustering, we could utilize the same ideas to estimate average treatment effects within each cluster and then average between clusters. For example, we could stratify, match, or weight observations within each cluster to yield an average cluster-level treatment effect. Averaging these cluster-level treatment effects would yield an overall estimate. This would solve some of the concerns about the lack of independence between observations.

In this dissertation we assumed that there was no clustering in the design or that its effect was negligible. This assumption that clustering at later stages of sampling can be ignored should be further investigated through additional simulation studies.

6.3.2 Missing Data

Missing data are often an issue of concern when it comes to sample surveys. For a propensity score analysis, several methods have been suggested for dealing with missing covariate data. Rosenbaum and Rubin propose a generalized propensity score [37], which is defined as the probability of treatment conditional on a set of observed covariates and a missing data pattern indicator. This method seeks to balance the missing data patterns along with the distributions of the covariates. Multiple imputation [40] is also suggested to deal with missing covariate data when we can assume
that the missing data mechanism is missing at random [14]. Both of these methods could be applied to the complex survey context.

When data are missing not at random, a different method is needed. One way of addressing the problem is from a sensitivity analysis perspective. Assume a 1:K matching design. Drawing on the work of Rosenbaum [34], we can assume different degrees of imbalance between the treated and control distributions of the covariates based on the missing data information. When all observations in a set are matched exactly on \( \hat{e} \), they should have equal probability of being the treated observation in the set; however, based on our assumptions about the degree of the imbalance because of missing data, we can characterize how much more likely one individual is to be treated rather than others in the same matched set, given a hypothetical unmeasured variable in the propensity score model. This allows us to estimate bounds on the p-value of a Wilcoxon signed-rank test. To implement this method with survey data, an analogous method to the Wilcoxon signed-rank test would need to developed for survey data or additional modeling assumptions would be needed.

6.3.3 Identifying Heterogeneity

This dissertation focuses on estimating the population average treatment effect when the treatment effect is possibly heterogeneous. A logical followup question is if we can identify the sources of heterogeneity. This could be done by looking at the estimated treatment effect as a function of the propensity score [47] or as a function of other covariates [29]. Translating and furthering these methods to the survey context would be an important extension to our research.
Bibliography


