Propensity Score Overlap Weighting

Ben Rohlfing

MS Mathematics, Statistics and Operations Research

Southern Illinois University Edwardsville

December 2019

1 Introduction

Estimating the treatment effect is a major goal in many observational studies. Direct comparison, however, of the target (response) variable between two treatment groups is generally discouraged because of possible imbalance in the background (covariate) variables between the two groups. For example, the THIN population-based cohort study [9] compared the mortality (response variable) of 30,811 statin users (treatment group) with 60,291 patients not using statins (control group). The THIN study covariates contained cardiovascular risk factors, age, sex, BMI, smoking, drinking, other medications, and other diseases (48 variables in all). Imbalance in the background covariates might lead to biased estimates of the treatment effect.

Propensity score methods attempt to estimate the causal effect of a treatment in observational studies by taking into account the covariates that would help predict whether a subject receives said treatment or not. The goal of this project is to give an overview of propensity score weighting methods and provide a detailed review of the results in a recently published paper by Li, F., Morgan, K. L., & Zaslavsky, A. M. (JASA, 2018) titled "Balancing

Covariates via Propensity Score Weighting."

The propensity score of an individual is the conditional probability of this individual receiving a treatment knowing this individual's pre-treatment or background variables (covariates). A treatment group is a group in an experiment that receives a certain treatment, while a control group is the group that doesn't receive said treatment while everything else from the treatment group remains the same, meaning it is the baseline group. For this type of study, let X be the $(1 \times p)$ vector of pre-treatment variables(covariates), Y is the observed outcome variable, and Z is the binary treatment variable.

The idea of using propensity score methods was first introduced in 1984 by Paul Rosenbaum and Donald Rubin in order to estimate causal effects in observational studies, and more specifically, the causal effects of smoking on one's mortality rate. The model used in this example is explained in a later section. Now a variety of propensity score weighting techniques are used across a wide variety of fields, such as economics, medical studies, psychological studies, and school performance given different types of post-secondary education institutions.

This paper will proceed as follows. Section 1 gives a general introduction to the problem of estimating the average treatment effect. Section 2 describes the Propensity Score Framework or Rubin Causal Model that practitioners use to formalize and provide a mathematical framework for causal inference. Section 3 introduces the necessary assumptions, definitions, and results which allow us to use propensity score methods. Section 4 introduces the logistic regression method which is used to estimate the propensity score of each individual unit. Section 5 introduces some of the intuition behind using propensity scores as weights in order to estimate treatment effects. Section 6 will explain some of the theory behind how covariates can be balanced using propensity score weighting. Section 7 will discuss properties of large-sample nonparametric estimators that will be used to estimate the average treatment effect. Section 8 will introduce overlap weighting, a type of propensity score weighting that is shown to yield consistent and efficient estimators. Section 9 will examine a real-life example to apply some of these concepts. In Section 10, sets of simulation experiments will be run that are designed to analyze the effectiveness of different propensity score weighting estimators.

2 Potential Outcome Framework

Suppose a sample of size n is obtained from some population where each unit in the sample can be categorized as belonging to one of two groups. Let $Z_i = z$ be the indicator variable identifying group membership with a control group ($z = 0$) and a treatment group ($z = 1$). For each unit i, let $X_i = (X_{i1}, \ldots, X_{iK})^T$ be the vector of K covariate measurements, and Y_i be the observed outcome or response variable - the variable that is in question in the study that changes with respect to the different values of the covariates.

The potential outcome framework [1] is widely used in causal inference where we define two types of potential outcomes: $Y_i(0)$ is the potential outcome for the ith individual if that individual is in the control group, and $Y_i(1)$ equals the potential outcome for the ith individual if that individual is in the treatment group. Using the potential outcomes framework, we can write the ith observed response as

$$Y_i = Y_i(Z_i) = Z_i \cdot Y_i(1) + (1 - Z_i) \cdot Y_i(0) \tag{1}$$

We are interested in the effect of the treatment on the ith individual. One common form of the individual treatment effect is $Y_i(1) - Y_i(0)$, which is the individual's potential outcome in the treatment group minus the potential outcome in the control group. Another commonly used individual treatment effect is the ratio of the potential outcomes, $Y_i(1)/Y_i(0)$. In this project we are only interested in the former.

The fundamental problem in causal inference is that the individual treatment effect, $Y_i(1) - Y_i(0)$, cannot be directly measured, since one person or unit cannot be in the treatment group and the control group at the same time. This means that one of the two values in this expression cannot be measured.

Instead of looking at the individual treatment effect, many researchers opt to estimate the average treatment effect (ATE) in the study population defined as

$$ATE = E[Y(1) - Y(0)]. \tag{2}$$

A naive estimator of the ATE is the difference of the sample average outcomes in the treatment group and the sample average outcomes in the control group,

$$\widehat{ATE}_{nv} = \frac{\sum_{i=1}^{n} Z_i Y_i}{\sum_{i=1}^{n} Z_i} - \frac{\sum_{i=1}^{n}(1 - Z_i)Y_i}{\sum_{i=1}^{n}(1 - Z_i)}. \tag{3}$$

3 Propensity Score Explained

This section will lay the foundation for explaining what a propensity score is by providing assumptions, definitions, and theorems that are foundational to the idea of propensity scores.

Assumption 3.1. *(Unconfoundedness) [2] For any unit $i = 1, \ldots, n$,*

$$P(Z_i = 1 \mid Y_i(0), Y_i(1), X_i) = P(Z_i = 1 \mid X_i). \tag{4}$$

or, using conditional independence notation

$$Z_i \perp\!\!\!\perp (Y_i(0), Y_i(1)) \mid X_i. \tag{5}$$

In other words, the treatment variable Z_i is independent of the potential outcomes, $Y_i(0)$ and $Y_i(1)$.

Assumption 3.1 above states that treatment assignment Z_i is conditionally independent of the potential outcomes $(Y_i(0), Y_i(1))$ when conditioned on the pre-treatment observed covariates. Thus, Assumption 3.1 says that the decision on whether or not the ith individual receives treatment does not depend on the potential outcome values after controlling for the pre-treatment variables.

Assumption 3.1 is necessary to avoid the problem of “self-selection bias”, which means that treatment is assigned to those who benefit most from it. For example, suppose there is a new drug ($Z_i = 1$) that improves the condition of a cancer patient (Y_i). However, this treatment is mostly administered to those and wanted by those that are in worse condition rather than being randomly assigned to patients. So, if we estimate the drug's effect, we would be comparing those patients who are in worse shape (treatment) to those who are in better shape (control). Thus, the ending result of the patient's outcome would be partly dependent on their prior condition and would thus lead to a biased result. However, Assumption 3.1 states that after controlling for the patient's background information, both those who took the drug and those who didn't would be equivalent in their remaining characteristics. Thus, the difference in patient condition is only attributed to the drug effect.

Assumption 3.2. *(Probabilistic Assignment or Positive Overlap) [2] For any unit i,*

$$0 < P(Z_i = 1 \mid X_i) < 1.$$

Definition 3.1. *(Balancing Score) [2] For every unit i in the sample, a balancing score $b(X_i)$ is a function of the covariate X_i such that*

$$Z_i \perp\!\!\!\perp X_i \mid b(X_i),$$

or, in terms of a probability statement,

$$P(Z_i = 1|\ X_i,\ b(X_i)) = P(Z_i = 1|\ b(X_i)).$$

Theorem 3.1. *(Unconfoundedness given any balancing score) [2] Suppose Assumption 1 is true. Then, treatment assignment is unconfounded given any balancing score,*

$$P(Z_i = 1 \mid Y_i(0), Y_i(1), b(X_i)) = P(Z_i = 1 \mid b(X_i)), \tag{6}$$

or, using conditional independence notation

$$Z_i \perp\!\!\!\perp (Y_i(0), Y_i(1)) \mid b(X_i)). \tag{7}$$

Proof. We start with the left hand side and show the right hand side of (6),

$$\begin{aligned} P(Z_i = 1 \mid Y_i(0), Y_i(1), b(X_i)) &= E_Z[Z_i \mid Y_i(0), Y_i(1), b(X_i)] \\ &= E_X[E_Z(Z_i \mid Y_i(0), Y_i(1), X_i, b(X_i)) \mid Y_i(0), Y_i(1), b(X_i)] \\ &= E_X[E_Z(Z_i \mid X_i, b(X_i)) \mid Y_i(0), Y_i(1), b(X_i)] \\ &\qquad \text{by Assumption 3.1.} \end{aligned}$$

Since conditioning on $b(X_i)$ in the outer expectation makes $E_Z(Z_i \mid X_i, b(X_i))$ in the inner expectation constant, we get

$$\begin{aligned} P(Z_i = 1 \mid Y_i(0), Y_i(1), b(X_i)) &= E_X[E_Z(Z_i \mid b(X_i)) \mid Y_i(0), Y_i(1), b(X_i)] \\ &= E_Z(Z_i \mid b(X_i)) \\ &= P(Z_i = 1 \mid b(X_i)). \end{aligned}$$

□

This proves that if the treatment variable is conditioned on a balancing score, unconfoundedness still holds.

Definition 3.2. *(Propensity Score) [1] The propensity score of unit i, with covariate measurement X_i, is defined as the conditional probability of treatment assignment*

$$e(X_i) = P(Z_i = 1|X_i) = E_Z[Z_i|X_i].$$

Theorem 3.2. *(Propensity Score is a balancing score) [2] For every unit i, the propensity score $e(X_i)$ is a balancing score, i.e.,*

$$P(Z_i = 1 \mid X_i, e(X_i)) = P(Z_i = 1 \mid e(X_i)). \tag{8}$$

Proof. Start with the left hand side of (8). Conditioning on X_i also conditions $e(X_i)$,

$$P(Z_i = 1 \mid X_i, e(X_i)) = P(Z_i = 1 \mid X_i) = e(X_i).$$

Now, with the right hand side of (8), using iterated expectations

$$\begin{aligned} P(Z_i = 1 \mid e(X_i)) &= E_Z[Z_i \mid e(X_i)] \\ &= E_X[E_Z[Z_i \mid X_i, e(X_i)] \mid e(X_i)] \\ &= E_X[E_Z[Z_i \mid X_i] \mid e(X_i)] \\ &= E_X[e(X_i) \mid e(X_i)] = e(X_i), \end{aligned}$$

again since conditioning on X_i also conditions $e(X_i)$. □

Theorem 3.2 shows that individuals from either treatment group with the same propensity score are "balanced" in a way that the distribution of X is the same regardless of the treatment group to which the individual belongs. In addition, Theorem 3.1 shows that the treatment group indicator is unrelated to the potential outcomes for individuals sharing the same propensity score.

4 Logistic Regression

The most common way to estimate propensity scores is to use logistic regression. Logistic regression is used to model a binary response variable ($Z_i = 0, 1$) using multiple predictors or covariates (X_i). By definition, the propensity score is

$$e(X_i) = P(Z_i = 1 \mid X_i) = E[Z_i \mid X_i],$$

since Z_i is a Bernoulli random variable with success probability $e(X_i)$. The binary logistic regression response function is [16]

$$e(X_i) = \frac{\exp(X_i^T \beta)}{1 + \exp(X_i^T \beta)}, \tag{9}$$

where $X_i \in \mathbb{R}^p$ is the vector of pre-treatment observed covariates for the i^{th} unit, and $\beta = (\beta_1, \ldots, \beta_p)^T$ is the vector of parameters. To estimate the response function, or propensity score $e(X_i)$, the vector of parameters, β, needs to be estimated using the commonly used method of "Maximum Likelihood."

The likelihood function of β can then be written as

$$L(\beta) = \prod_{i=1}^{n} e(X_i)^{Z_i} \cdot (1 - e(X_i))^{1-Z_i}.$$

Thus, the log-likelihood function for β under the logistic regression model is

$$\begin{aligned} l(\beta) = \ln(L(\beta)) &= \ln\Big(\prod_{i=1}^{n} e(X_i)^{Z_i} \cdot (1 - e(X_i))^{1-Z_i}\Big) \\ &= \sum_{i=1}^{n} \ln\Big(e(X_i)^{Z_i} \cdot (1 - e(X_i))^{1-Z_i}\Big) \\ &= \sum_{i=1}^{n} \Big(Z_i \cdot \ln(e(X_i)) + (1 - Z_i) \cdot \ln(1 - e(X_i))\Big). \qquad (10) \end{aligned}$$

To estimate the MLE's for β, we must first get the partial derivative of the log-likelihood function, ℓ, with respect to β

$$\frac{\partial l}{\partial \beta} = \sum_{i=1}^{n} \Big(\frac{Z_i}{e(X_i)} \cdot e'(X_i) - \frac{1 - Z_i}{1 - e(X_i)} \cdot (1 - e'(X_i))\Big). \qquad (11)$$

Note that $e(X_i)$ is a function of β, so $e'(X_i)$ is short for $\partial e(X_i)/\partial\beta$. Next, the derivative $\frac{\partial l}{\partial \beta}$ is equated to 0 and solved for $\hat{\beta}_1 = b_1, ..., \hat{\beta}_p = b_p$, which maximize the log-likelihood function. We rely on numerical methods to solve these equations.

Consequently, the fitted logistic response function is

$$\hat{e}(X_i) = \frac{\exp(X_i^T b)}{1 + \exp(X_i^T b)}, \text{ where } b = (b_1, \ldots, b_p)^T.$$

The fitted response values are then the estimated propensity scores. All these computations can be done routinely using R software [15].

5 Intuition of Weighted Estimators

The majority of this project analyzes the idea of using weighted estimators, with the weights being the propensity scores or functions of propensity scores.

In order to explain this concept, a simple example will be used. Figure 1 is an example of a distribution of propensity scores for the control units (blue line) and the treatment units (red line).

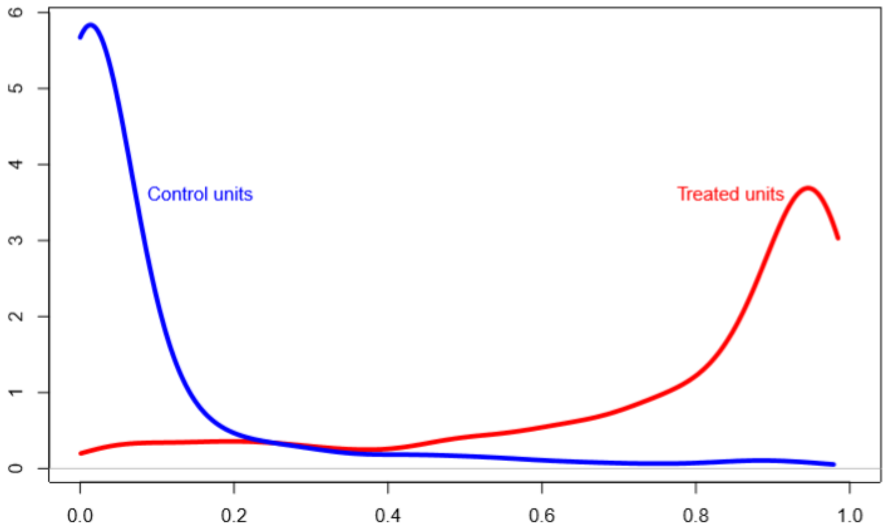

Figure 1: Propensity Score Distribution

One of the main goals of using propensity score methods is to find any similarities between the treatment and control group distributions. However, as noted in the image, the distributions are not very similar. One way to remedy this is to weight the treated and control distributions using the propensity scores.

If the ATE is being estimated (ATE is the estimand), a treatment unit is weighted by the reciprocal of the propensity score of the unit $e(X_i)$, while a control unit is weighted by the compliment of the propensity score of the unit $1-e(X_i)$. In other words, the treatment units with low propensity scores and control units with high propensity scores are upweighted. The theory behind this intuition is given below. We can write $E[Z \cdot Y]$ as

$$\begin{aligned} E[Z \cdot Y] &= E[Z \cdot Y(1)] \qquad \text{(Equation (1))} \\ &= E_X(E[Z \cdot Y(1)|X]) \qquad \text{(iterated expectation)} \\ &= E_X(E[Z|X] \cdot E[Y(1)|X]) \quad \text{(Assumption 3.1: } Y(1) \perp\!\!\!\perp Z|X) \\ &= E_X(e(X) \cdot E[Y(1)|X]) \quad \text{(Definition 3.2).} \end{aligned}$$

By Assumption 3.2, $0 < e(X) < 1$, we can rewrite $E[Y(1)]$ as

$$E\left\{\frac{Z \cdot Y}{e(X)}\right\} = E_X(E[Y(1)|X]) = E[Y(1)].$$

An unbiased estimator for $E[Y(1)]$, assuming $e(X_i)$ is known or can be estimated, is

$$\widehat{E[Y(1)]} = \frac{1}{n}\sum_{i=1}^{n} \frac{Z_i Y_i}{e(X_i)}. \tag{12}$$

In a similar argument, we can show that

$$E\left\{\frac{(1-Z)\cdot Y}{1-e(X)}\right\} = E[Y(0)],$$

which leads to its unbiased estimator

$$\widehat{E[Y(0)]} = \frac{1}{n}\sum_{i=1}^{n} \frac{(1-Z)Y}{1-e(X)}. \tag{13}$$

Finally, we can write the ATE as

$$E[Y(1)-Y(0)] = E[Y(1)] - E[Y(0)] = E\left\{\frac{Z\cdot Y}{e(X)}\right\} - E\left\{\frac{(1-Z)\cdot Y}{1-e(X)}\right\}. \tag{14}$$

Consequently, an unbiased estimator of the ATE (14), based on (12) and (13), is

$$\begin{aligned}\widehat{ATE}_w &= \widehat{E[Y(1)]} - \widehat{E[Y(0)]} \\ &= \frac{1}{n}\sum_{i=1}^{n} \frac{Z_i Y_i}{e(X_i)} - \frac{1}{n}\sum_{i=1}^{n} \frac{(1-Z_i)Y_i}{1-e(X_i)}\end{aligned} \tag{15}$$

If the average treatment effect of the treated units, or the ATT, is being estimated (ATT is the estimand), only the control unit is weighted by the ratio of its propensity score and its compliment. We start with

$$E[Y(1)-Y(0)|Z_i = 1] = E[Y(1)|Z=1] - E[Y(0)|Z=1]$$

Note that we can write $E[Y(1)|Z=1] = E[Z\cdot Y(1)] = E[Z\cdot Y]$ with an unbiased estimator of

$$E[Y\widehat{(1)|Z} = 1] = \frac{1}{n}\sum_{i=1}^{n} Z_i Y_i. \tag{16}$$

Now, we can write $E[Y(0)|Z=1]$ as

$$\begin{aligned}
E[Y(0)|Z=1] = E[Z \cdot Y(0)] &= E_X\big[E[Z \cdot Y(0)|X]\big] \qquad \text{(Equation 1)} \\
&= E_X\big[E[Z|X] \cdot E[Y(0)|X]\big] \quad \text{(Assumption 3.1)} \\
&= E_X\left[\frac{e(X)}{1-e(X)}(1-E[Z|X]) \cdot E[Y(0)|X]\right] \quad \text{(Definition 3.2)} \\
&= E_X\left[\frac{e(X)}{1-e(X)} \cdot E[(1-Z)Y(0)|X]\right] \quad \text{(Assumption 3.1)} \\
&= E_X\left\{E\left[\frac{e(X)}{1-e(X)} \cdot (1-Z)Y \middle| X\right]\right\} \quad \text{(Equation 1)} \\
&= E\left[\frac{e(X)}{1-e(X)} \cdot (1-Z)Y\right].
\end{aligned}$$

Thus, an unbiased estimator of $E[Y(0)|Z=1]$ is

$$E[Y\widehat{(0)|Z}=1] = \frac{1}{n}\sum_{i=1}^{n} \frac{(1-Z_i)Y_i e(X_i)}{1-e(X_i)}. \tag{17}$$

Consequently, the ATT can be written as

$$\begin{aligned}
E[Y(1)-Y(0)|Z=1] = E[Y(1)|Z=1] - E[Y(0)|Z=0] \\
= E[ZY] - E\left[\frac{(1-Z)Ye(X)}{1-e(X)}\right].
\end{aligned} \tag{18}$$

Finally, a natural estimator of the ATT (18), assuming that $e(X_i)$ is known or can be estimated, based on (16) and (17), is

$$\begin{aligned}
\widehat{ATT}_w &= E[Y\widehat{(1)|Z}=1] - E[Y\widehat{(0)|Z}=1] \\
&= \frac{1}{n}\sum_{i=1}^{n} Z_i Y_i - \frac{1}{n}\sum_{i=1}^{n} \frac{(1-Z_i)Y_i \cdot e(X_i)}{1-e(X_i)}.
\end{aligned} \tag{19}$$

6 Balancing Covariates via Propensity Score Weighting

Let X be the vector of covariates with PDF $f(X_i)$. Fan et al. (2018) introduced the idea of conditional average controlled difference (ACD) which is defined as

$$\begin{aligned}\tau(x) &= E[Y(1) - Y(0)|X = x]\\ &= E[Y|Z = 1, X = x] - E[Y|Z = 0, X = x].\end{aligned}$$

The target population density is defined as $f(x)h(x)$, where $h(x)$ is the weight function of x. Various forms of the weight function $h(x)$ are covered in the current section and Section 8.

First, we look at a discrete covariate case, for example, where X only takes three values, $X = 1, 2, 3$. Let $\tau(1)$, $\tau(2)$, and $\tau(3)$ be the conditional average control difference for $X = 1, 2, 3$. Then, the expected ACD and ATE of $\tau(x)$ over the target population can be defined as

$$\tau_h = \frac{\tau(1)P(X = 1)h(1) + \tau(2)P(X = 2)h(2) + \tau(3)P(X = 3)h(3)}{P(X = 1)h(1) + P(X = 2)h(2) + P(X = 3)h(3)}.$$

When X is continuous or a vector of continuous measurements, the weighted ACD over the target population is defined as

$$\tau_h = \frac{\int \tau(x)f(x)h(x)dx}{\int f(x)h(x)dx}. \tag{20}$$

Let $f_z(x) = f(X = x|Z = z)$. Note that

$$\begin{aligned}f_1(x) = f(X = x|Z = 1) &= \frac{f(X = x, Z = 1)}{P(Z = 1)}\\ &= \frac{f(X = x) \cdot P(Z = 1|X = x)}{P(Z = 1)} = \frac{f(x) \cdot e(x)}{P(Z = 1)}.\end{aligned}$$

Thus, we can write $f_1(x) \propto f(x) \cdot e(x)$ and, similarly, $f_0(x) \propto f(x) \cdot (1 - e(x))$.

This concept can be extended to the target population density, where

$$f(x)h(x) \propto \frac{f_1(x)}{e(x)} \cdot h(x) = f_1(x)\omega_1(x)$$

is true for $Z = 1$, while

$$f(x)h(x) \propto \frac{f_0(x)}{1 - e(x)} \cdot h(x) = f_0(x)\omega_0(x)$$

is true for $Z = 0$. The ω's are defined as the balancing weights of the target population density, where

$$\omega_1(x) = \frac{h(x)}{e(x)},\ \omega_0(x) = \frac{h(x)}{(1 - e(x))}. \tag{21}$$

The function $h(x)$ can be set to anything, depending on what is to be estimated. For example, in order to estimate the ATE, set $h(x) = 1$. The weights (ω_1, ω_0) then become $(1/e(x), 1/(1 - e(x)))$. If the ATT is to be estimated, then set $h(x) = e(x)$, where the weights become $(1, e(x)/(1 - e(x)))$. For the ATC estimation, set $h(x) = 1 - e(x)$, with corresponding weights $((1 - e(x))/e(x), 1)$.[10]

Finally, let us define the Weighted Average Treatment Effect (WATE) as

$$\hat{\tau}_h = \frac{\sum_i \omega_1(x_i)Z_iY_i}{\sum_i \omega_1(x_i)Z_i} - \frac{\sum_i \omega_0(x_i)(1 - Z_i)Y_i}{\sum_i \omega_0(x_i)(1 - Z_i)}. \tag{22}$$

7 Large-Sample Nonparametric Estimator Properties

Definition 7.1. *Consistency. Let $Z_1, Z_2, ...$ be iid random variables and Z be a random variable. The random variable Z_n converges in probability to Z, or $Z_n \xrightarrow{p} Z$ if*

$$\lim_{x \to \infty} P\big(|Z_n - Z| \le \epsilon\big) = 1$$

.

Definition 7.2. *Consistency of an estimator. Let $X_1, X_2, ...$ be a sequence of iid random variables drawn from a distribution with parameter θ and $\hat{\theta}$ as its estimator. This estimator $\hat{\theta}$ is a consistent esitmator of θ if*

$$\hat{\theta} \xrightarrow{p} \theta \quad or \quad \lim_{n \to \infty} P\big(|\hat{\theta}(X_1, ..., X_n) - \theta| \le \epsilon\big) = 1.$$

Lemma 7.1. *Strong Law of Large Numbers: The sample average converges almost surely to* *the expected value,*

$$\bar{X}_n \xrightarrow{a.s.} \mu, \textit{ as } n \to \infty, \quad \textit{or} \quad P\left(\lim_{n\to\infty} \bar{X}_n = \mu\right) = 1.$$

Lemma 7.2. *(Slutsky's Theorem) Let $\hat{\theta} \xrightarrow{p} \theta$ and $\hat{\eta} \xrightarrow{p} \eta$. Then, for any continuous multivariate valued function g,*

$$g(\hat{\theta}, \hat{\eta}) \xrightarrow{p} g(\theta, \eta).$$

Theorem 7.1. *[10] $\hat{\tau}_h$ is a consistent estimator of τ_h.*

Here we are not proving the unbiasedness of $\hat{\tau}_h$, $E[\hat{\tau}_h] = \tau_h$. We are proving the consistency of the estimator, which means that as $n \to \infty$, $\hat{\tau}_h \to \tau_h$ in probability as defined in Definition 7.2.

Proof. Without loss of generalization, let's consider the continuous version of the ACD. The categorical version will be similar except we will be using the probability mass function and the integral will be replaced with a summation. Recall the ACD defined as

$$\tau_h = \frac{\int \tau(x) f(x) h(x) dx}{\int f(x) h(x) dx}.$$

Let's start by rewriting $\tau(x)$ as

$$\begin{aligned}
\tau(x) &= E_{Y,Z|X}[Y(1)-Y(0)|X=x] = E_{Y,Z|X}[Y(1)-Y(0)|Z,X=x] \quad \text{(Assumption 3.1)}\\
&= E_{Y,Z|X}[Y(1)|Z,X=x] - E_{Y,Z|X}[Y(0)|Z,X=x]\\
&= E_{Y,Z|X}[Y|Z=1,X=x] - E_{Y,Z|X}[Y|Z=0,X=x]\\
&= E_{Y,Z|X}[ZY|Z=1,X=x] - E_{Y,Z|X}[(1-Z)Y|Z=0,X=x]\\
&= \int\int zyf(y|Z=1,x)dydz - \int\int (1-z)yf(y|z=0,x)dydz\\
&= \int\int \frac{zyf(y,z=1,x)dydz}{f(z=1,x)} - \int\int \frac{(1-z)yf(y,z=0,x)dydz}{f(z=0,x)}\\
&= \int\int \frac{zyf(x)f(y,z=1|x)dydz}{f(x)P(z=1|x)} - \int\int \frac{(1-z)yf(x)f(y,z=0|x)dydz}{f(x)P(z=0|x)}\\
&= \int\int \frac{zyf(y,z=1|x)dydz}{e(x)} - \int\int \frac{(1-z)yf(y,z=0|x)dydz}{1-e(x)}\\
&= E_{Y,Z|X}\Big[\frac{Z\cdot Y}{e(x)}\Big|X=x\Big] - E_{Y,Z|X}\Big[\frac{(1-Z)\cdot Y}{1-e(x)}\Big|X=x\Big],
\end{aligned}$$

where $e(x) = P(Z=1|X=x)$ is the propensity score. Now this is substituted back into the original integral:

$$\begin{aligned}
\int \tau(x)f(x)h(x)dx &= \int \left(E_{Y,Z|X}\Big[\frac{Z\cdot Y}{e(x)}\Big|x\Big] - E_{Y,Z|X}\Big[\frac{(1-Z)\cdot Y}{1-e(x)}\Big|x\Big]\right) h(x)f(x)dx\\
&= \int \left(E_{Y,Z|X}\Big[\frac{h(x)}{e(x)}\cdot ZY\Big|x\Big] - E_{Y,Z|X}\Big[\frac{h(x)}{1-e(x)}\cdot(1-Z)Y\Big|x\Big]\right) f(x)dx.
\end{aligned}$$

With the top part of the ACD completed, the focus now shifts to the bottom. Here, we focus on rewriting $h(x)$ into a piecewise function as

$$h(x) = \begin{cases} z\cdot\frac{h(x)}{e(x)}P(z=1|x), & z=1\\ (1-z)\cdot\frac{h(x)}{1-e(x)}P(z=0|x), & z=0. \end{cases}$$

Since $Z|x$ is a Bernoulli random variable with success probability $P(z=1|x)$, we can write $h(x)$ in terms of conditional expectations as

$$h(x) = \begin{cases} E_{Z|X}\Big[\frac{h(x)}{e(x)}\cdot Z\Big|x\Big], & z=1\\ E_{Z|X}\Big[\frac{h(x)}{1-e(x)}\cdot(1-Z)\Big|x\Big], & z=0. \end{cases}$$

Putting together the top and bottom yields

$$\begin{aligned}
\tau_h &= \frac{\int \left\{ E_{Y,Z|X}\left[\frac{h(x)}{e(x)} \cdot ZY\Big|x\right] - E_{Y,Z|X}\left[\frac{h(x)}{1-e(x)} \cdot (1-Z)Y\Big|x\right]\right\} f(x)dx}{\int f(x)h(x)dx} \\
&= \frac{\int E_{Y,Z|X}\left[\frac{h(x)}{e(x)} \cdot ZY\Big|x\right] f(x)dx}{\int f(x)h(x)dx} - \frac{\int E_{Y,Z|X}\left[\frac{h(x)}{1-e(x)} \cdot (1-Z)Y\Big|x\right] f(x)dx}{\int f(x)h(x)dx} \\
&= \frac{\int E_{Y,Z|X}\left[\frac{h(x)}{e(x)} \cdot ZY\Big|x\right] f(x)dx}{\int E_{Z|X}\left[\frac{h(x)}{e(x)} \cdot Z\Big|x\right] f(x)dx} - \frac{\int E_{Y,Z|X}\left[\frac{h(x)}{1-e(x)} \cdot (1-Z)Y\Big|x\right] f(x)dx}{\int E_{Z|X}\left[\frac{h(x)}{1-e(x)} \cdot (1-Z)\Big|x\right] f(x)dx} \\
&= \frac{\int E_{Y,Z|X}\left[\omega_1(x) \cdot ZY\Big|x\right] f(x)dx}{\int E_{Z|X}\left[\omega_1(x) \cdot Z\Big|x\right] f(x)dx} - \frac{\int E_{Y,Z|X}\left[\omega_0(x) \cdot (1-Z)Y\Big|x\right] f(x)dx}{\int E_{Z|X}\left[\omega_0(x) \cdot (1-Z)\Big|x\right] f(x)dx}. \qquad (23)
\end{aligned}$$

Recall the WATE estimator

$$\hat{\tau}_h = \frac{\frac{1}{n}\sum_i \omega_1(x_i)Z_iY_i}{\frac{1}{n}\sum_i \omega_1(x_i)Z_i} - \frac{\frac{1}{n}\sum_i \omega_0(x_i)(1-Z_i)Y_i}{\frac{1}{n}\sum_i \omega_0(x_i)(1-Z_i)}. \qquad (24)$$

Note that each component in the WATE estimator (24) converges to each component in (23) by Law of Large Numbers. Finally, by Slutsky's Theorem the WATE estimator $\hat{\tau}_h$ will converge in probability to τ_h. □

Next, let us consider the variance of the estimator $\hat{\tau}_h$. Let $\mathbf{X} = \{x_1, ..., x_n\}$ represent the sampled covariate design points. Using iterated expectations, we can write

$$\begin{aligned}
Var(\hat{\tau}_h) = E(\hat{\tau}_h^2) - [E(\hat{\tau}_h)]^2 &= E_X(E[\hat{\tau}_h^2|\mathbf{X}]) - (E_X(E[\hat{\tau}_h|\mathbf{X}]))^2 \\
&= E_X\big[Var(\hat{\tau}_h|\mathbf{X}) + (E[\hat{\tau}_h|\mathbf{X}])^2\big] - (E_X[E[\hat{\tau}_h|\mathbf{X}]])^2 \\
&= E_X[Var(\hat{\tau}_h|\mathbf{X})] + E_X([E[\hat{\tau}_h|\mathbf{X}])^2 - (E_X[E[\hat{\tau}_h|\mathbf{X}]])^2 \\
&= E_X[Var(\hat{\tau}_h|\mathbf{X})] + Var_X(E[\hat{\tau}_h|\mathbf{X}]), \qquad (25)
\end{aligned}$$

where the first term is the expected variation directly due to variation in $\mathbf{X}$, while the second term is the unexplained variation coming from somewhere other than $\mathbf{X}$. The second term in (25), $Var_X(E[\hat{\tau}_h|\mathbf{X}])$, can be attributed to the dependence of the expectations in the WATE estimator (24) on the sample, and estimating it involves the outcome model (i.e. model between the potential outcomes $Y(z)$ and the covariate x) [10]. Also, Imbens (2004) argued that the first term in (25), $E_X[Var(\hat{\tau}_h|\mathbf{X})]$, is typically much larger than $Var_X(E[\hat{\tau}_h|\mathbf{X}])$. So, the benefit of selecting weights that incorporate the outcome model would not justify the risk of biasing the model specification to attain the results [10, 14]. Hence, we focus our attention on the selection of weights $h(x)$ that will minimize $E_X[Var(\hat{\tau}_h|\mathbf{X})]$. The next theorem shows the asymptotic properties of $E_X[Var(\hat{\tau}_h|\mathbf{X})]$.

Theorem 7.2. *[10] As $n \to \infty$, the expectation of the conditional variance of the estimator $\hat{\tau}_h$, given the sample $\mathbf{X} = \{x_1, ..., x_n\}$ converges:*

$$n \cdot E_X(Var[\hat{\tau}_h|\mathbf{X}]) \to \frac{\int f(x)h(x)^2[v_1(x)/e(x) + v_0(x)/1 - e(x))]dx}{\int [h(x)f(x)]^2 dx},$$

where $v_z(x) = Var[Y(z)|\mathbf{X}]$.

Proof. Now, consider first the conditional variance in the first term. If we further condition on $\mathbf{Z} = \{z_1, \ldots, z_n\}$, we have

$$\begin{aligned} Var(\hat{\tau}_h|\mathbf{X}, \mathbf{Z}) &= \frac{\sum_{i=1}^n \omega_1^2(x_i)Z_i^2 \cdot Var(Y_i|x_i, Z_i)}{[\sum_{i=1}^n \omega_1^2(x_i)Z_i]^2} + \frac{\sum_{i=1}^n \omega_0^2(x_i)(1 - Z_i)^2 \cdot Var(Y_i|x_i, Z_i)}{[\sum_{i=1}^n \omega_0^2(x_i)(1 - Z_i)]^2} \\ &= \frac{\sum_{i=1}^n \frac{(h(x_i))^2}{(e(x_i))^2} Z_i \cdot Var(Y_i(1)|x_i)}{[\sum_{i=1}^n \frac{h(x_i)}{e(x_i)} Z_i]^2} + \frac{\sum_{i=1}^n \frac{(h(x_i))^2}{(1-e(x_i))^2}(1 - Z_i) \cdot Var(Y_i(0)|x_i)}{[\sum_{i=1}^n \frac{h(x_i)}{1-e(x_i)}(1 - Z_i)]^2} \\ &= \frac{\frac{1}{n}\sum_{i=1}^n \frac{Z_i}{e(x_i)}[(h(x_i))^2/e(x_i)] \cdot v_1(x_i)}{n[\frac{1}{n}\sum_{i=1}^n \frac{Z_i}{e(x_i)} h(x_i)]^2} + \frac{\frac{1}{n}\sum_{i=1}^n \frac{1-Z_i}{1-e(x_i)}[(h(x_i))^2/(1 - e(x_i))] \cdot v_0(x_i)}{n[\frac{1}{n}\sum_{i=1}^n \frac{1-Z_i}{1-e(x_i)} h(x_i)]^2}. \end{aligned}$$

Note that $v_z(x_i) = Var(Y_i(Z_i)|x_i)$. Also note that

$$E_Z\Big[\frac{Z_i}{e(x_i)}|x_i\Big] = \frac{1}{e(x_i)} E_Z[Z_i|x_i] = \frac{1}{e(x_i)} \cdot e(x_i) = 1 \tag{26}$$

and

$$E_Z\Big[\frac{1 - Z_i}{1 - e(x_i)}|x_i\Big] = \frac{1}{1 - e(x_i)} E_Z[1 - Z_i|x_i] = \frac{1}{1 - e(x_i)} \cdot (1 - e(x_i)) = 1. \tag{27}$$

Hence the sample version of (26) and (27) will both approach 1 by Strong of Law of Large Numbers. Next, let's average the $Var(\hat{\tau}_h|\mathbf{X}, \mathbf{Z})$ over the distribution of $\mathbf{Z}$ and using the equalities (26) and (27) together with Slutsky's Theorem, we attain

$$n \cdot E_X Var(\hat{\tau}_h|\mathbf{X}) = n \cdot E_X E_Z[Var(\hat{\tau}_h|\mathbf{X}, \mathbf{Z})] \longrightarrow \frac{\int \Big[\frac{v_1(x)}{e(x)} + \frac{v_0(x)}{1-e(x)}\Big] \cdot h(x)^2 f(x)dx}{\big(\int h(x)f(x)dx\big)^2}.$$

If the residual variance is then assumed to be homoscedastic across both groups, say $v_1(x) = v_0(x) = v$, then the asymptotic variance of $\hat{\tau}_h$ simplifies to

$$n \cdot E_X Var[\hat{\tau}_h|\mathbf{X}] \to v \cdot \frac{\int f(x)h(x)^2/[e(x)(1-e(x))]dx}{\left(\int h(x)f(x)dx\right)^2}.$$

□

Lemma 7.3. *(Cauchy-Schwarz Inequality) If X and Y are random variables, then*

$$[E(XY)]^2 \leq E(X^2)E(Y^2).$$

Proof. Consider $W = (X - aY)^2$, with a being a constant. Since $W \geq 0$, we have

$$\begin{aligned} 0 \leq E(W) &= E(X - aY)^2 \\ &= E(X^2) - 2aE(XY) + a^2E(Y^2). \end{aligned} \tag{28}$$

Since a is any constant, let $a = \frac{E(XY)}{E(Y^2)}$. Substituting this into (28) leads to

$$\begin{aligned} 0 &\leq E(X^2) - 2\frac{\left[E(XY)\right]^2}{E(Y^2)} + \left[\frac{E(XY)}{E(Y^2)}\right]^2 E(Y^2) \\ &= E(X^2) - 2\frac{\left[E(XY)\right]^2}{E(Y^2)} + \frac{\left[E(XY)\right]^2}{E(Y^2)} \\ &= E(X^2) - \frac{\left[E(XY)\right]^2}{E(Y^2)} \\ &= \frac{E(X^2)E(Y^2) - \left[E(XY)\right]^2}{E(Y^2)}. \end{aligned}$$

Rearranging the above inequality yields

$$\left[E(XY)\right]^2 \leq E(X^2)E(Y^2).$$

□

The following corollary establishes the important asymptotic result that justifies the use of the overlap weights, $h(x) = e(x)(1 - e(x))$, in the WATE estimator.

Corollary 7.1. *[10] The function $h(x) \propto e(x)(1-e(x))$ gives the smallest asymptotic variance for the weighted estimator $\hat{\tau}_h$ among all h's under homoscedasticity, and as $n \to \infty$,*

$$n \cdot \min_h (E_X Var[\hat{\tau}_h|\mathbf{X}]) \to \frac{v}{C_h^2} \cdot \int f(x)e(x)(1-e(x))dx,$$

where $C_h = \int h(x)f(x)dx$.

Proof. Recall the result of Lemma 7.3 where, given that X and Y are random variables,

$$[E(XY)]^2 \leq E(X^2)E(Y^2).$$

Thus, we have

$$\big(E[h(x)]\big)^2 = \left[E\left(\frac{h(x)}{\sqrt{e(x)(1-e(x)}}\sqrt{e(x)(1-e(x)}\right)\right]^2 \leq E\left[\frac{h(x)^2}{e(x)(1-e(x)}\right]E[e(x)(1-e(x)].$$

Equality is attained when

$$\frac{h(x)}{\sqrt{e(x)(1-e(x)}} \propto \sqrt{e(x)(1-e(x))} \implies h(x) \propto e(x)(1-e(x)).$$

When applying this property to the right hand side of Theorem 7.2, we get

$$\begin{aligned} n \cdot E_X Var[\hat{\tau}_h|\mathbf{X}] &\to \frac{v}{C_h^2} \cdot E\left[\frac{h(x)^2}{e(x)(1-e(x)}\right] \\ &\geq \frac{v}{C_h^2} \cdot \frac{\big(E[h(x)]\big)^2}{E[e(x)(1-e(x)]} \\ &= \frac{v}{C_h^2} \cdot E[e(x)(1-e(x)], \end{aligned}$$

where the last equality is true when $h(x) = e(x)(1-e(x))$. □

8 Overlap Weighting

The overlap weights can be defined as $h(x) = e(x)(1-e(x))$, or they can be defined as

$$h(x_i) = \begin{cases} \omega_1(x_i) \propto 1 - e(x_i) & z = 1 \\ \omega_0(x_i) \propto e(x_i) & z = 0 \end{cases}$$

Among all weights, the corresponding nonparametric estimator $\hat{\tau}_h$ has a minimum asymptotic variance. This concept is illustrated by Figure 2.

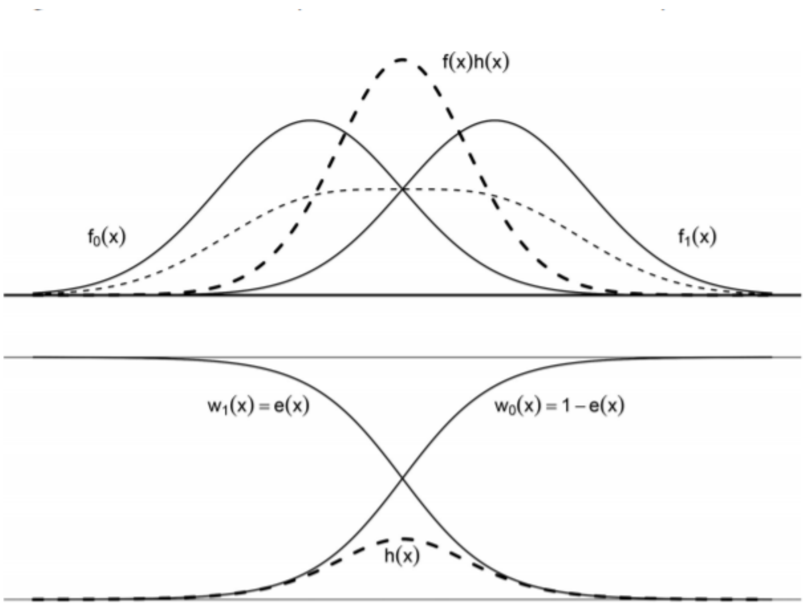

Figure 2: Overlap Weights Using Two Normal Distributions

Here, we use the example of two univariate normal distributions with different means but equal variance. The dashed curve indicates the target population density $f(x_i)h(x_i)$ in which the treatment and control groups most overlap, hence where the corresponding WATE is called ATO (average treatment effect of the overlap population).

The lower panel dashed curve gives the overlap distribution $h(x_i)$ where the distribution falls under both normal distributions. Here, the propensity scores fall near 0.5, meaning individual observations here could be in either group. These observations would then be weighted more heavily. Also, since the overlap distribution is relatively bounded, it is less susceptible to extreme weights, thus minimizing $\hat{\tau}_h$'s asymptotic variance.

A big advantage of this variance-minimizing property of the overlap weights is that it can adapt to any distribution of covariates and propensities. For example, for any small propensity to treatment, ATO approximates ATT (for $e(x_i) \approx 0, (1 - e(x_i), e(x_i)) \approx \left(1, \frac{e(x_i)}{1-e(x_i)}\right)$. Now, if propensity to control is small, ATO appoximates ATC. For a dataset where the treatment and control groups are nearly balanced in size and distribution, ATO approximates

ATE (for $e(x_i) \approx 0.5, (1-e(x_i), e(x_i)) \approx \left(\frac{.25}{e(x_i)}, \frac{.25}{1-e(x_i)}\right)\Big)$.

This overlap distribution and its population represents an area where the individual observations of interest could realistically appear in either the treatment or control group, which makes this an area of increased analysis of observational studies. For example, if a medicine is being administered, it is important to gauge its true effects by looking at the people with similar characteristics where treatment assignment can go either way instead of comparing treatment effects on two different groups of people.

The next theorem shows that when the propensity score is estimated using a logistic regression model, then the overlap weights guarantee exact balance in the weighted covariates.

Theorem 8.1. *[10] When the propensity scores are estimated by maximum likelihood under a logistic regression model, the overlap weights lead to exact balance in the means of any included covariate between treatment and control groups. In other words,*

$$\frac{\sum_i x_{ik} Z_i (1-\hat{e}(x_i))}{\sum_i Z_i(1-\hat{e}(x_i))} = \frac{\sum_i x_{ik}(1-Z_i)\hat{e}(x_i)}{\sum_i (1-Z_i)\hat{e}(x_i)}, \quad k = 1, ..., K.$$

Proof. We start by defining the score functions of the logistic propensity score model, where $\text{logit}[e(x_i)] = \beta_0 + x_i\beta'$, with $\beta = (\beta_1, ..., \beta_k)$.

$$\frac{\partial \log L}{\partial \beta_x} = \sum x_{ik}(Z_i - \hat{e}(x_i)) = 0,$$

where $x_{0k} = 1$ and $\hat{e}(x_i) = [1 + exp(-x_i\beta')]^{-1}$.

If $x_{0k} = 1$, then $\sum(Z_i - \hat{e}(x_i)) = 0$ and $\sum Z_i = \sum \hat{e}(x_i)$. More generally, $\sum x_{ik}Z_i = \sum x_{ik}\hat{e}(x_i)$.

It follows that

$$\sum Z_i(1-\hat{e}(x_i)) = \sum Z_i - \sum Z_i\hat{e}(x_i) = \sum \hat{e}(x_i) - \sum Z_i\hat{e}(x_i) = \sum \hat{e}(x_i)(1-Z_i)$$

and

$$\begin{aligned}\sum x_{ik}Z_i(1-\hat{e}(x_i)) &= \sum x_{ik}Z_i - \sum x_{ik}Z_i\hat{e}(x_i)\\ &= \sum x_{ik}\hat{e}(x_i) - \sum x_{ik}Z_i\hat{e}(x_i) = \sum \hat{x}_{ik}e(x_i)(1-Z_i).\end{aligned}$$

Thus, for any $k = 1, ..., K$,

$$\frac{\sum x_{ik}Z_i(1-\hat{e}(x_i))}{\sum Z_i(1-\hat{e}(x_i))} = \frac{\sum x_{ik}Z_i(1-\hat{e}(x_i))}{\sum \hat{x}_{ik}e(x_i)(1-Z_i)}.$$

□

9 Example: Right Heart Catheterization

Right heart catheterization (RHC) is "a diagnostic procedure for directly measuring cardiac function in critically ill patients." This example will look at the effectiveness of RHC through the use of different propensity score weights.

The dataset of interest is available publicly. The study contains data on 5735 adult patients, 2184 of which underwent the RHC procedure ($Z = 1$), while the remaining 3551 patients didn't receive the procedure ($Z = 0$). The outcome was a binary variable `dth30` which measured whether or not a patient survived after 30 days of admission ($Y = 1$ if they died, $Y = 0$ if not).

The dataset included 72 covariates for analyzing. There were 21 continuous covariates, 25 binary covariates, and 26 dummy variables. These dummy variables were formed by breaking up 6 categorical covariates by using the `fastDummies` package in R.

The propensity score was then estimated under a logistic model with the 72 covariates against the treatment variable `swang1`. The covariate mean balance is then measured using the absolute standardized bias (ASB), which is given by the following equation:

$$ASB = \left| \frac{\sum_{i=1}^{N} x_i Z_i \omega_i}{\sum_{i=1}^{N} Z_i \omega_i} - \frac{\sum_{i=1}^{N} x_i (1-Z_i)\omega_i}{\sum_{i=1}^{N}(1-Z_i)\omega_i} \right| \Bigg/ \sqrt{s_1^2/N_1 + s_0^2/N_0}.$$

Here, s_z^2 is the variance of the unweighted covariate of interest for treatment group z, while N_z is the sample size of each treatment group z. The weights ω_i are based on which estimate is being used to analyze the balance. For the ASB measurements, the unweighted covariates (in which case the weights would be removed from the equation), the ATE, ATT, and the overlap weights will be applied and compared to each other to see which has the best covariate balance (lower ASB values mean better balance).

In order to calculate this, some modifications to the data were performed, especially with consideration to the variance of each treatment group for each covariate. To make this easier, a new assignment variable z was created that changed the treatment variable `swang1` from "No" to 0 and "Yes" = 1. This way, the ASB could be calculated with these numbers, and the data could be reordered such that the assignment groups were together, which allowed for much easier variance calculations of the two groups.

Another dataset was created that eliminated some of the unused covariates in the dataset. The remaining factor covariates were turned into integer covariates for the purposes of calculating the ASB.

Now, for each weight type, the denominators and weights were calculated for each treatment group. For loops were then created to calculate the ASB for each covariate and stored in a vector to be used to create boxplots of the ASB distribution for each weight type. These boxplots are presented in Figure 3.

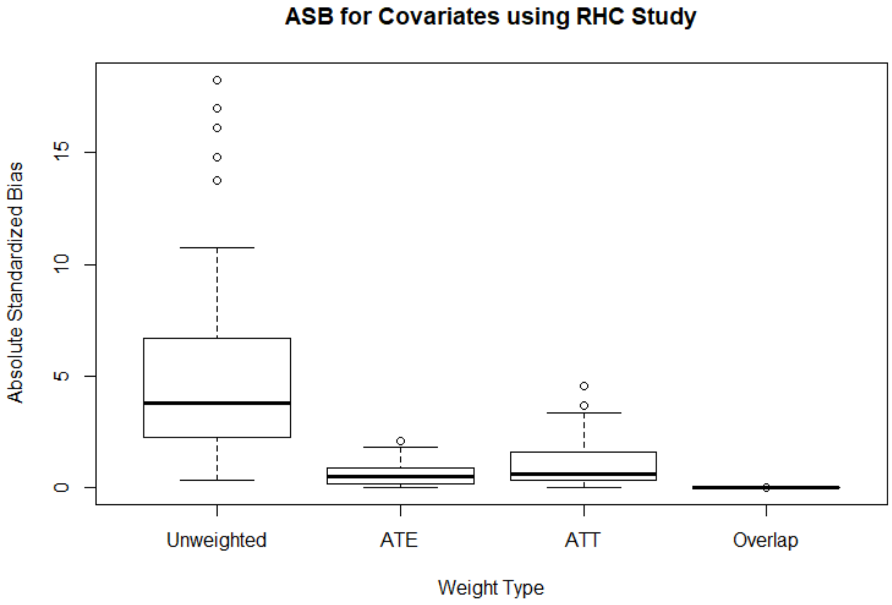

Figure 3: ASB Boxplot Comparison of Different Weights

As clearly seen, the unweighted data is the least balanced by a lot, while the ATE and ATT improve balance markedly. The overlap weights, though, clearly produce the best covariate balance, as the ASB for each covariate is basically zero.

The next step is to estimate the WATE using each weight. The following WATE equation was used:

$$\widehat{WATE} = \frac{\sum_i \omega_1(x_i) Z_i Y_i}{\sum_i \omega_1(x_i) Z_i} - \frac{\sum_i \omega_0(x_i)(1 - Z_i) Y_i}{\sum_i \omega_0(x_i)(1 - Z_i)}.$$

Before this could happen, the outcome variable `dth30` was transformed into an integer variable from a factor by turning "No" and "Yes" into 0 and 1, respectively, for the purposes of calculation.

Besides the other weights, a truncated ATT version of the WATE was calculated. This was done by putting the data in ascending order based on propensity score and cutting off the data points that had propensity scores less that 0.1 or greater than 0.9.

The standard errors obtained from each method were also calculated using simple bootstrapping techniques. A function was created that first takes a random sample of the dataset

Table 1: Treatment Effect Weighted Estimates and Standard Errors

	Unweighted	ATE	ATT	Overlap	Trunc. ATT
Estimate $\cdot 10^2$	7.36	5.50	5.40	6.41	5.77
SE $\cdot 10^2$	1.39	1.82	2.36	1.47	1.67

indices, applies them to a new dataset, and then creates a logistic model to estimate the propensity score based on this new dataset. The weights and denominators for each method were then created using another function, the WATE's were calculated using yet another function inside the bootstrapping function and then returned. This was replicated 100 times using the `replicate` function in R, from which the standard errors for each method were calculated.

The treatment effect weighted estimates and standard errors are given in Table 1. Note that the WATE estimator with overlap weights produced the smallest standard error among all weighted estimators.

10 Simulation

A simulation involving a wide array of propensity score balancing methods was run in order to test the effectiveness of each method.

The design of this simulation is as follows. First, six variables $V_1 - V_6$, are generated from a multivariate normal distribution with a mean of zero, a sample size of $n = 500$, and a correlation of 0.5 between each variable. Thus, the covariance matrix is compound symmetric, with a value of 1 for the diagonal entries and a value of 0.5 for every other entry. Next, $V_1 - V_3$ were kept as continuous covariates $X_1 - X_3$ and $V_4 - V_6$ were dichotomized such that negative values turned into 1 and positive values turned into zero. These binary variables were called $X_4 - X_6$.

Next, the propensity scores for each observation were calculated using a logistic model as follows:

$$e(X_n) = (1 + exp[-(\alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \alpha_3 X_3 + \alpha 4 X_4 + \alpha_5 X_5 + \alpha_6 X_6)])^{-1}.$$

Here, the parameters in the propensity score model are

$$(\alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5, \alpha_6) = (0.15\gamma, 0.3\gamma, 0.3\gamma, -0.2\gamma, -0.25\gamma, -0.25\gamma).$$

The γ values range from 1 to 4, where 1 represents higher overlap between the treatment and control groups and weak tails for each distribution, while 4 represents low overlap between the groups and strong tails for each distribution. Also, α_0 represents the overall treatment prevalence in each sample. This value can be either 0.4 or 0.1 in this simulation.

Next, each observation is assigned to either the treatment or control group by simulating a Bernoulli model given each observations' propensity score. This column in each dataset will be represented as Z.

The outcome variable Y is then calculated as follows:

$$E[Y|Z, X] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \Delta Z.$$

In this formula, $\Delta = 0.75$ to reflect the beneficial effect of the treatment.

Now, the various propensity score weighting methods that were used are described here. Note that all estimates follow the WATE formula given at the end of Section 6.

First, a crude estimate of the WATE is calculated, where the weights $\omega_i = 1$. This is the baseline estimate, and is used to compare with the other methods.

Next, the overlap weighting (OW) method WATE is calculated using the propensity scores, where $\omega_1 = 1 - \hat{e}(x_i)$ if $Z_i = 1$ and $\omega_0 = \hat{e}(x_i)$ if $Z_i = 0$. As discussed in Section 8, the individuals with propensity scores around 0.5 represent the highest proportion of individuals and thus would be the most important individuals to analyze, and eliminate extreme cases that have very little to do with the analysis at hand.

The other method discussed for WATE estimation is inverse probability weighting (IPW). As discussed earlier, $\omega_1 = \frac{1}{\hat{e}(x_i)}$ if $Z_i = 1$ and $\omega_0 = \frac{1}{1-\hat{e}(x_i)}$ if $Z_i = 0$. This ensures that extreme

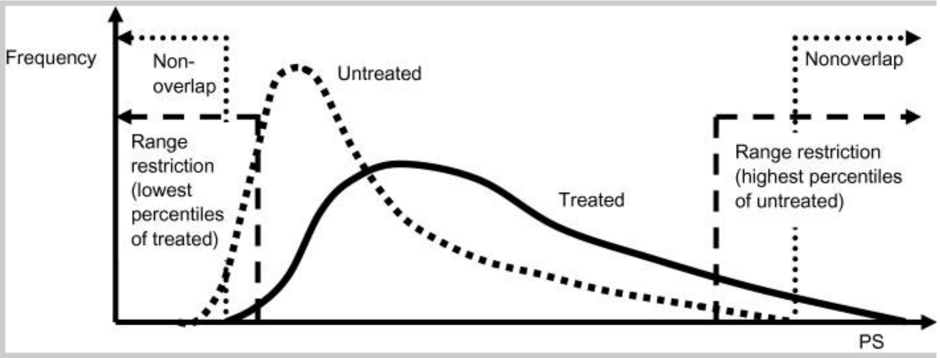

Figure 4: Asymmetric Trimming Overview[12]

values are given enough weight in order to bring the treatment and control groups under similar distributions.

Three different methods with inverse probability weighting, or IPW, are analyzed. The weights used for the IPW are the same weights used for the weighted ATE. First, the regular, untrimmed datasets were used to calculate the IPW.

Next, symmetric trimming in which each individual whose propensity score is outside the range $[\alpha, 1-\alpha]$ is cut out of the dataset and the IPW weights are recalculated using the trimmed dataset. In this simulation, three different alpha levels, $\alpha = 0.05, \alpha = 0.10$, and $\alpha = 0.15$, are analyzed.

Finally, an asymmetric trimming method in which multiple steps are involved is also analyzed. The first step involves removing individuals outside of the overlap of propensity scores between the treatment and control groups. In other words, the control units with a PS lower than the lowest treatment unit PS are eliminated, along with the treatment units with a higher PS than the highest control unit PS. After this step is performed, the treatment units with a PS below the q quantile of the treated units and the control units with a PS above the $(1-q)$ quantile of all the control units are removed. If $q = 0$, then only the first step is performed. The IPW is then calculated for each individual in the asymmetrically trimmed dataset. An overview of this method is given in Figure 4.

This gives a total of nine different WATE estimator methods to be analyzed: the crude

Table 2: Bias of Estimators with Treatment Prevalence = 0.4

Estimator	$\gamma = 1$	$\gamma = 2$	$\gamma = 3$	$\gamma = 4$
Crude	-2.00	-3.18	-3.77	-4.08
Overlap Weighting	0.00	-0.01	-0.02	-0.02
IPW				
No trimming	0.00	-0.04	-0.23	-0.54
Symmetric trimming				
α=0.05	0.00	-0.04	-0.05	-0.03
α=0.10	0.00	-0.02	-0.02	-0.04
α=0.15	-0.01	-0.01	-0.02	-0.02
Asymmetric trimming				
$q = 0$	0.18	0.44	0.74	0.90
$q = 0.01$	-0.25	-0.47	-0.54	-0.56
$q = 0.05$	-1.03	-1.55	-1.69	-1.60

estimate, the overlap weighting estimate, the untrimmed IPW estimate, the symmetrically trimmed IPW estimates with $\alpha = 0.05, \alpha = 0.10$, and $\alpha = 0.15$, and the asymmetrically trimmed IPW with $q = 0, q = 0.01$,and $q = 0.05$. Each of these methods are analyzed at each of the four γ levels ($\gamma = 1, 2, 3, 4$) and each of the two treatment prevalence values ($\alpha_0 = 0.1, 0.4$), giving a total of 72 method variations to analyze. For each of these method variations, 1000 data sets are simulated and used to calculate the desired values.

The first thing to be analyzed is the bias of each estimator. This is done by taking the mean of the 1000 estimators from the 1000 replicates and subtracting the treatment effect, $\Delta = 0.75$, from that. The results for a treatment prevalence of 0.4 are given in Table 2 and a treatment prevalence of 0.1 in Table 3.

The crude estimator is heavily biased, especially with increasing levels of γ. This is true for all of the estimators, that with increasing levels of γ, or increasingly stronger-tailed distributions, that the bias increases. This bias is miniscule with the overlap weighting and symmetrically trimmed IPW, with higher bias for untrimmed IPW and the asymmetrically trimmed IPW with increasing levels of q.

Table 3: Bias of Estimators with Treatment Prevalence = 0.1

Estimator	$\gamma = 1$	$\gamma = 2$	$\gamma = 3$	$\gamma = 4$
Crude	-2.01	-3.19	-3.78	-4.09
Overlap Weighting	0.00	-0.01	-0.02	-0.02
IPW				
No trimming	-0.01	-0.05	-0.23	-0.58
Symmetric trimming				
$\alpha = 0.05$	-0.01	-0.04	-0.05	-0.03
$\alpha = 0.10$	-0.01	-0.02	-0.03	-0.03
$\alpha = 0.15$	-0.01	-0.01	-0.03	-0.02
Asymmetric trimming				
$q = 0$	0.18	0.46	0.77	0.91
$q = 0.01$	-0.25	-0.41	-0.54	-0.56
$q = 0.05$	-1.03	-1.53	-1.68	-1.57

The next step is to calculate the estimate of the root mean square error (RMSE) of the 1000 estimators based on the 1000 replicates for each method. The formula for the estimated RMSE of each method is given by

$$RMSE(\hat{\theta}) = \sqrt{Var(\hat{\theta}) + \left[E[\hat{\theta}] - \theta\right]^2}.$$

This is calculated for each method, and the results for the treatment prevalence levels of 0.4 and 0.1 are given in Tables 4 and 5, respectively.

These results closely mirror those of the bias estimates. The crude estimators have a high RMSE estimate that gets higher with increasing γ. The untrimmed IPW estimators also follow this pattern with lower estimated RMSE's than the crude ones. The asymmetrically trimmed IPW estimators exhibit higher estimated RMSE's with increasing γ and q. The overlap and symmetrically trimmed IPW estimators show very little variation in RMSE estimate values from different γ and α values.

The last metric to be calculated was the 95 percent normality-based 95 percent confidence interval coverage. This was done by generating 100 different data sets, bootstrapping each data set 100 times to find the mean and variance of any given estimator in each data set,

Table 4: RMSE of Estimators with Treatment Prevalence = 0.4

Estimator	$\gamma = 1$	$\gamma = 2$	$\gamma = 3$	$\gamma = 4$
Crude	2.02	3.19	3.78	4.08
Overlap Weighting	0.29	0.29	0.30	0.32
IPW				
No trimming	0.35	0.60	0.97	1.36
Symmetric trimming				
$\alpha = 0.05$	0.35	0.41	0.42	0.40
$\alpha = 0.10$	0.35	0.33	0.33	0.34
$\alpha = 0.15$	0.32	0.30	0.30	0.32
Asymmetric trimming				
$q = 0$	0.36	0.65	1.02	1.31
$q = 0.01$	0.41	0.62	0.73	0.76
$q = 0.05$	1.08	1.60	1.76	1.68

and then figure out if the confidence interval of the estimator for each data set contains 0.75. Only the crude, overlap, untrimmed IPW, and $\alpha = 0.15$ trimmed IPW estimators were run for time constraints and was only used to confirm certain results. The results are given in Table 6.

The analysis of 95 percent confidence interval results confirms some of the prior results. The crude estimators have no interval coverage, meaning they are far off from the treatment estimate. This is not a surprise, considering the high bias and RMSE of all the crude estimators. However, the overlap weighting estimators have sufficient interval coverage, meaning the overlap weights give a sufficient estimate of the treatment effect. This is once again not a surprise, given its low bias and RMSE. The same logic can be applied to the $\alpha = 0.15$ trimmed IPW estimator interval coverage. Finally, when looking at the untrimmed IPW estimator, the interval coverage decreases with increasing values of γ, which is consistent with the bias and RMSE results. [11]

Table 5: RMSE of Estimators with Treatment Prevalence = 0.1

Estimator	$\gamma = 1$	$\gamma = 2$	$\gamma = 3$	$\gamma = 4$
Crude	2.03	3.20	3.79	4.10
Overlap Weighting	0.29	0.30	0.30	0.30
IPW				
No trimming	0.35	0.64	1.03	1.41
Symmetric trimming				
$\alpha = 0.05$	0.35	0.43	0.41	0.38
$\alpha = 0.10$	0.35	0.34	0.32	0.34
$\alpha = 0.15$	0.33	0.30	0.30	0.33
Asymmetric trimming				
$q = 0$	0.37	0.69	1.08	1.33
$q = 0.01$	0.41	0.58	0.73	0.77
$q = 0.05$	1.09	1.59	1.75	1.65

Table 6: 95 Percent CI Coverage of Estimators with Treatment Prevalence = 0.4

Estimator	γ=1	γ=2	$\gamma = 3$	$\gamma = 4$
Crude	0.00	0.00	0.00	0.00
Overlap Weighting	0.96	0.97	0.94	0.94
IPW				
No trimming	0.92	0.87	0.40	0.02
Symmetric trimming				
$\alpha = 0.15$	0.94	0.97	0.95	0.90

References

[1] Rosenbaum, Paul R., and Donald B. Rubin. “The central role of the propensity score in observational studies for causal effects.” Biometrika 70.1 (1983): 41-55.

[2] Rosenbaum, Paul R., and Donald B. Rubin. “Reducing Bias in Observational Studies Using Subclassification on the Propensity Score.” Journal of the American Statistical Association 79, no. 387 (September 1984): 193-206.

[3] Rubin, Donald B. ”Using propensity scores to help design observational studies: appli-

cation to the tobacco litigation." Health Services and Outcomes Research Methodology 2.3-4 (2001): 169-188.

[4] Dietrich, Cecile C., and Eric J. Lichtenberger. "Using Propensity Score Matching to Test the Community College Penalty Assumption." The Review of Higher Education 38, no. 2 (Winter 2015): 193-219.

[5] Helmreich, James E., and Robert M. Pruzek. "PSAgraphics: AnRPackage to Support Propensity Score Analysis." Journal of Statistical Software 29, no. 6 (2009).

[6] Stampf, Susanne. "Propensity Score Based Data Analysis. "March 28, 2014, 1-30.

[7] Stuart, Elizabeth A., Gary King, Kosuke Imai, and D. E. Ho. "MatchIt: nonparametric preprocessing for parametric causal inference." Journal of Statistical Software 42.8 (2011).

[8] Craycroft, John. "Propensity Score Methods : A Simulation and Case Study Involving Breast Cancer Patients." Electronic Theses and Dissertations.

[9] Heart Protection Study Collaborative Group. "MRC/BHF Heart Protection Study of cholesterol lowering with simvastatin in 20 536 high-risk individuals: a randomised placebo controlled trial." The Lancet 360.9326 (2002): 7-22.

[10] Li, Fan, Kari Lock Morgan, and Alan M. Zaslavsky. "Balancing covariates via propensity score weighting." Journal of the American Statistical Association 113.521 (2018): 390-400.

[11] Li, Fan, Laine E. Thomas, and Fan Li. "Addressing Extreme Propensity Scores via the Overlap Weights." American Journal of Epidemiology 188, no. 1 (September 5, 2018): 250–57.

[12] Stürmer, Til, Kenneth J. Rothman, Jerry Avorn, and Robert J. Glynn. “Treatment Effects in the Presence of Unmeasured Confounding: Dealing With Observations in the Tails of the Propensity Score Distribution—A Simulation Study.” American Journal of Epidemiology 172, no. 7 (August 17, 2010): 843–54.

[13] Imai, Kosuke, and Marc Ratkovic. “Covariate Balancing Propensity Score.” Journal of the Royal Statistical Society 76, no. 1 (March 2013): 243–63.

[14] Imbens, Guido W. “Nonparametric estimation of average treatment effects under exogeneity: A review.” Review of Economics and statistics 86.1 (2004): 4-29.

[15] R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/. (2018)

[16] Kutner, Michael H., Chris Nachtsheim, and John Neter. Applied linear regression models. McGraw-Hill/Irwin (2004).