An Overview of Treatment Effect Estimation in the Potential Outcomes Framework

Thierry Rioual *supervised by* Pierre-Henri Wuillemin

LIP6 - Sorbonne University

August 9, 2024

Abstract

A treatment effect refers to the causal effect of a given treatment on an outcome variable of interest. In the Neyman-Rubin potential outcomes framework, a treatment effect is defined for each individual unit in terms of two potential outcomes. Each unit has one outcome that would result from receiving the treatment and another outcome that would result from not receiving the treatment. The treatment effect is therefore the difference between these two potential outcomes.

In this paper, we will provide an overview of different cases of treatment effect estimation and present relevant estimators. We will cover the estimation of average treatment effects in experimental studies, observational studies, and also discuss the estimation of heterogeneous treatment effects.

1. Potential Outcomes

1.1 *The Neyman-Rubin Causal Model*

Consider a set of n units from a given population, each associated with data describing their features. A certain treatment is assigned to a subset of the population, while the remaining units do not receive the treatment. Following the treatment assignment, a specific quantity of interest, referred to as the outcome, is selected. The difference in this outcome between the treated and untreated groups, known as the *treatment effect*, is measured to capture the causal effect of the treatment on the outcome. This treatment effect is the central focus of the potential outcomes model.

Formally, denote the data consisting of the units' features, treatment assignment, and observed outcome as an *observation*. The framework models the observations using a family of n random variables $(O_i)_{i\in[\![1,n]\!]}$, each having the same structure as $O = (X, T, Y)$ where:

- $X = (X^{(1)}, \dots, X^{(d)}) \in \mathbb{R}^d$ is a covariate vector consisting of the features of a subject.
- $T \in \{0, 1\}$ is the treatment assignment with $T = 0$ indicating that a subject received the *control treatment* (i.e. did not receive the treatment), and $T = 1$ indicating that it received the *active treatment* (i.e. did receive the treatment).
- $Y \in \mathbb{R}$ is the *observed* outcome of the treatment on a subject, with $Y(1)$ and $Y(0)$ representing the *potential* treated and untreated outcomes, respectively.

We aim to quantify the causal effect of a given treatment. For the i-th subject in the studied population, this is given by the *Individual Treatment Effect* (ITE):

$$\Delta_i = Y_i(1) - Y_i(0)$$

However, it is impossible to see both potential outcomes at once, since one of the potential outcomes is always missing. This dilemma is known as *the fundamental problem of causal inference*, rendering direct estimation from the data infeasible. The potential outcomes that are not (and cannot) be observed are referred to as *counter factuals*, as they counter the actual, factual outcome observed in reality of the studied units. †

Although the per-unit treatment effect Δ is fundamentally unknowable, statistical estimations from data enable us to infer certain properties about it. In practice, we will primarily focus on estimating the *Average Treatment Effect* (ATE): †

$$\tau = \mathbb{E}[Y(1) - Y(0)]$$

When we condition the expectation on the covariates X, we obtain the *Conditional Average Treatment Effect* (CATE), given by:

$$\tau(x) = \mathbb{E}[Y(1) - Y(0) \mid X = x]$$

This quantity is particularly useful when studying heterogeneous treatment effects, especially in populations that respond differently to treatment. In such cases, the ATE might be close to zero, while significant CATE values could exist for specific covariate values, indicating substantial treatment effects for certain subgroups.

1.2 *Assumptions*

To effectively estimate the (unobservable) ATE, it is essential to make certain assumptions regarding the variables to draw meaningful conclusions.

1.a. *Ignorability/Exchangeability*

The ignorability assumption assures that the treatment assignment for a subject is independent of their potential outcomes. In other words, this implies that a subject's likelihood of receiving the treatment is not influenced by how they would potentially respond to it.

$$T \perp\!\!\!\perp \{Y(0), Y(1)\}$$

It is important to distinguish the expression $T \perp\!\!\!\perp \{Y(0), Y(1)\}$ from $T \perp\!\!\!\perp Y$ as the former denotes that the treatment is independent of the *potential* outcomes, whereas the latter denotes independence for the *observed* outcome. In general, $T \not\perp\!\!\!\perp Y$, unless the treatment has no effect on the patient.

1.b. *Uncofoundedness/Conditional Exchangeability*

In many real-world applications, the ignorability assumption is not likely to hold, as units with certain treatment outcomes might have differing probabilities of receiving

† One might consider administering the control treatment first, followed by the active treatment, to study the treatment effect. However, this approach is not feasible, as the time difference between the two treatments would cause changes in the unit's features, thereby invalidating the treatment effect we aim to study. We are specifically interested in estimating the hypothetical scenario of "what if" the subject had been exposed to the alternative treatment under the exact same time period and conditions. This scenario is inherently unobservable.

† The ATE is also refered to as the "average causal effect" (ACE).

the treatment. To address this issue, a weaker and more plausible assumption, known as uncofoundedness, can be employed instead:

$$T \perp\!\!\!\perp \{Y(0), Y(1)\} \mid X$$

This assumption allows for the estimation of treatment effects in observational studies, even when the treatment T is influenced by the covariates X, provided that conditional independence is maintained.

2. *Stable Unit Treatment Value Assumption (SUTVA)*

   The Stable Unit Treatment Value Assumption is verified if a unit's outcome is imply a function of its treatment, it is equivalent to the following conditions:

   - *No Interference Between Units*
     The potential outcome for any unit does not depend on the treatment assigned to other units.
     $$\forall i \in [\![1, n]\!] \quad Y_i(T_1, \ldots, T_i, \ldots, T_n) = Y_i(T_i)$$
   - *Consistency*:
     The observed outcome for a unit under treatment is the same as the potential outcome for that unit under that treatment.
     $$\forall i \in [\![1, n]\!] \quad Y_i = Y_i(T_i) = T_i Y_i(1) + (1 - T_i) Y_i(0)$$

   Note that this may be a reasonable assumption in medical practices (i.e., the treatment prescribed to patient A does not affect patient B), however, it is less appropriate in social or economic settings where network effects may arise.

3. *Positivity/Overlap*

   Positivity is the condition that all subgroups of the data, defined by different covariate values, have a non-negligible probability of receiving each possible treatment.

   $$\forall t \in \{0, 1\}, \forall x \in \mathbb{R}^d \quad 0 < \mathbb{P}[T = t \mid X = x] < 1$$

   This assumption is crucial for computing the Average Treatment Effect (ATE), as it ensures that we do not condition on an event with zero probability.

An interesting phenomenon arising in setting up the study population is the Positivity-Unconfoundedness Tradeoff. As we condition on more covariates to help satisfy unconfoundedness, it could also lead to positivity violation. Indeed, by increasing the dimension of the covariates, we form smaller subgroups which have a higher chance of only having one type of treatment. [†] [4]

Using the aforementioned assumptions, we can effectively derive the estimand from observational data:

$$\begin{aligned}
\tau &= \mathbb{E}[Y(1) - Y(0)] = \mathbb{E}_X\left[\mathbb{E}[Y(1) \mid X] - \mathbb{E}[Y(0) \mid X]\right] \\
&= \mathbb{E}_X\left[\mathbb{E}[Y(1) \mid T = 1, X] - \mathbb{E}[Y(0) \mid T = 0, X]\right] && \text{(Uncofoundness and Positivity)} \\
&= \mathbb{E}_X\left[\mathbb{E}[Y \mid T = 1, X] - \mathbb{E}[Y \mid T = 0, X]\right] && \text{(STUVA)}
\end{aligned}$$

[†] For example, if the size of a subgroup is one, positivity is not satisfiable since the single subject only received one treatment, rendering the other treatment a zero probability.

This equality is sometimes referred to as the adjustment formula and provides the identifiability of the ATE. We will subsequently explore how to obtain an actual estimate from this formula.

1.3 *Randomized Controlled Trials*

A primary approach to estimating Average Treatment Effects involves the use of Randomized Controlled Trials (RCTs), which are experimental study designs widely regarded as the gold standard for evaluating the efficacy of interventions or treatments. RCTs employ randomization to help satisfy the ignorability assumption, which ensures that the potential outcomes are independent of the treatment assignment. In practice, an RCT is conducted by randomly allocating participants among the compared treatments. This randomization provides statistical control over various influences. The random assignment of participants to treatments reduces both selection bias and allocation bias, balancing both known and unknown prognostic factors in the treatment assignment.

2. First Approaches

2.1 *Difference in Means*

In a RCT environment, a naive first approach involves computing the mean observed outcomes of the subjects who received the treatment and subtracting it to the mean outcome of those who did not:

$$\hat{\tau}_{DM} = \frac{1}{n_1} \sum_{i \in \mathcal{T}_1} Y_i - \frac{1}{n_0} \sum_{i \in \mathcal{T}_0} Y_i$$

Where $\mathcal{T}_t = \{i \in [\![1, n]\!] \mid T_i = t\}$ and $n_t = |\mathcal{T}_t|$, with $t \in \{0, 1\}$. Using previous assumptions, we obtain an asymptotically normal, unbiased estimator of τ. [5]

The Difference in Means estimator is one of the simplest estimators of the Average Treatment Effect. However, it is valid only in RCT scenarios, as it relies on the ignorability assumption. When ignorability is not assumed but only unconfoundedness is, we can propose a variation of the DM estimator by aggregating over the covariates.

2.2 *Aggregated Difference in Means*

Assume that the potential outcomes for each unit are conditionally independent of its treatment assignment given its features. Let the covariate vectors X_i take values in a finite discrete space $\mathcal{X}$, where $|\mathcal{X}| \ll n$. Under positivity, for each covariate state $x \in \mathcal{X}$, there exists at least one unit receiving the active treatment and at least one unit receiving the control treatment.

Let $\mathcal{P} = \{[x_1], \ldots, [x_k]\}$ be a partition of the set $[\![1, n]\!]$ based on the equivalence relation $\sim$, where $i \sim j \Leftrightarrow X_i = X_j$, and let $\text{Rep}(\mathcal{R})$ denote the representative set for the equivalence classes of $\mathcal{P}$. Denote the equivalence classes with their respective cardinalities as

$$[x] = \{i \in [\![1, n]\!] \mid i \sim x\}, \; n_x = |[x]|$$

Then, let $n_{x,t} = |[x] \cap \mathcal{T}_t|$ for $t \in \{0, 1\}$.

Since we are not conditionning on a zero probability we can use the previous estimator to estimate the CATE:

$$\hat{\tau}_{DM}(x) = \frac{1}{n_{x,1}} \sum_{i \in [x] \cap \mathcal{T}_1} Y_i - \frac{1}{n_{x,0}} \sum_{i \in [x] \cap \mathcal{T}_0} Y_i$$

Under these assumption, we can estimate the ATE by aggregating the CATE estimations:

$$\hat{\tau}_{Agg} = \sum_{x \in \text{Rep}(\mathcal{P})} \frac{n_x}{n} \hat{\tau}_{DM}(x) = \sum_{x \in \text{Rep}(\mathcal{P})} \frac{n_x}{n} \left(\frac{1}{n_{x,1}} \sum_{i \in [x] \cap \mathcal{T}_1} Y_i - \frac{1}{n_{x,0}} \sum_{i \in [x] \cap \mathcal{T}_0} Y_i \right)$$

This initial approach is promising but encounters difficulties when the treatment distribution is unbalanced or when the potential outcomes exhibit high variance. We can see that each observation either contributes to improving the estimation of $\mathbb{E}[Y(0)]$ or $\mathbb{E}[Y(1)]$, but not both simultaneously. The rationale behind the following estimators will be to utilize the available data to address the gaps introduced by the fundamental problem of causal inference.

3. Learner based estimators

3.1 *S-Learners*

A first approach to linear-based methods for estimating treatment effects is derived from modeling the outcome as a function of the treatment assignment $T = t \in \{0, 1\}$ and the covariates $X = x \in \mathbb{R}^d$:

$$\mu(t, x) = \mathbb{E}[Y \mid T = t, X = x]$$

In this framework, statistical models, such as linear regression can be used to predict the counterfactual outcomes that are not directly observed in a Difference-in-Means estimation. Here, the ATE can be expressed as:

$$\tau = \mathbb{E}_X[\mu(1, X) - \mu(0, X)]$$

By denoting $\hat{\mu}$ as the estimation of the conditional expectation function μ, we obtain an estimator of the CATE given by:

$$\hat{\tau}_S(x) = \hat{\mu}(1, x) - \hat{\mu}(0, x)$$

Which in turn gives us the ATE:

$$\hat{\tau}_S = \frac{1}{n} \sum_{i=1}^{n} \hat{\mu}(1, X_i) - \hat{\mu}(0, X_i)$$

This approach is commonly referred to as the S-learner, where "S" signifies "Single", indicating the use of a unified model to determine the outcome. S-learners aims to improve the estimation of treatment effects when the treatment is unevenly distributed among the units, by generating artificial data points based on the statistical models. †

However, S-Learners may suffer from bias, as high-dimensional covariates can dilute the treatment effect, potentially skewing the ATE towards zero. To address this potential bias, T-learners employ separate models to fit the outcomes for treated and untreated units, thereby allowing for greater heterogeneity in the resulting estimates.

† A broader class of methods used to estimate treatment effects by modeling potential outcomes conditional on covariates and treatment assignments is referred to as Conditional Outcome Modeling (COM). S-learners are a specific instance of this broader class, utilizing a single unified function to model the outcome based on both covariates and treatment assignments. [4]

3.2 *T-Learners*

By defining the conditional expectations of the two *potential* outcomes as functions of the covariates:

$$\mu_0(x) = \mathbb{E}[Y(0) \mid X = x]$$
$$\mu_1(x) = \mathbb{E}[Y(1) \mid X = x]$$

We can apply similar techniques to predict μ_0 and μ_1. Under this definition, the ATE is expressed as:

$$\tau = \mathbb{E}_X[\mu_1(X) - \mu_0(X)]$$

An estimator of the ATE can be obtained by computing the empirical average of the differences between the predicted functions evaluated at the covariates for each unit. Let $\hat{\mu}_0$ and $\hat{\mu}_1$ denote the estimates of μ_0 and μ_1, respectively, derived from the models (which do not necessarily need to be the same). Estimator of the CATE and ATE are therefore given by:

$$\hat{\tau}_T(x) = \hat{\mu}_1(x) - \hat{\mu}_0(x)$$
$$\hat{\tau}_T = \frac{1}{n}\sum_{i=1}^{n} \hat{\mu}_1(X_i) - \hat{\mu}_0(X_i)$$

This approach is known as the "T-learner", where "T" denotes "Two", reflecting the use of separate models for each group. †

The T-learner approach aims to address the limitations of high-dimensional covariates in treatment effect estimation that are not adequately managed by S-learners. However, this method may result in fewer samples available for each model, which can be particularly problematic when treatment groups are unevenly distributed, potentially leading to one group having very few units.

3.3 *X-Learners*

To address the limitations of the S-learner and T-learner approaches, the X-learner leverages information from the control group to improve estimations for the treatment group, and vice versa. [1]

The X-learner starts similarly to the T-learner by estimating the conditional outcome expectation functions μ_0 and μ_1 through statistical models, yielding $\hat{\mu}_0$ and $\hat{\mu}_1$, respectively. However, instead of directly estimating the ATE, the X-learner calculates the ITE Δ_i by using the predictions in the counterfactual terms. This results in two sets of ITEs: one for the control group, $\hat{\Delta}_i^{(0)}$, and one for the treatment group, $\hat{\Delta}_i^{(1)}$. For $i \in [\![1, n]\!]$:

$$\hat{\Delta}_i^{(0)} = \hat{\mu}_1(X_i) - Y_i \text{ , if } T_i = 0$$
$$\hat{\Delta}_i^{(1)} = Y_i - \hat{\mu}_0(X_i) \text{ , if } T_i = 1$$

† T-learners fall into the broader class of Grouped Conditional Outcome Modeling (GCOM) estimators. GCOM extends the Conditional Outcome Modeling (COM) framework by incorporating a grouping mechanism to improve the estimation of treatment effects

Note that if $\hat{\mu}_0 = \mu_0$ and $\hat{\mu}_1 = \mu_1$, then $\tau(x) = \mathbb{E}[\hat{\Delta}_i^{(0)} \mid X = x] = \mathbb{E}[\hat{\Delta}_i^{(1)} \mid X = x]$. Once we have the two sets of ITEs, a supervised learning method is employed to estimate $\tau(x)$ from the control group data, giving us $\hat{\tau}_0(x)$, and similarly from the treatment group data, yielding $\hat{\tau}_1(x)$.

Finally, we the CATE estimator is defined as the weighted average of the two estimators:

$$\hat{\tau}_X(x) = g(x)\hat{\tau}_0(x) + (1 - g(x))\hat{\tau}_1(x)$$

where $g(x) \in [0, 1]$ is a weight function, often chosen to be the propensity score $e(x) = \mathbb{P}[T = 1 \mid X = x]$, which will be discussed in the next section.

The X-learner has two key advantages over other CATE estimators. First, it can adapt to structural properties such as the sparsity or smoothness of the CATE. This adaptability is particularly useful since the CATE is often zero or approximately linear. Second, the X-learner is especially effective when the number of units in one treatment group (usually the control group) is much larger than in the other.

3.4 *Ordinary Least Squares example*

A particular case of the previous Learner-based methods arises when we further assume a linear relationship between the outcome variable and the covariates, such that for all $i \in [\![1, n]\!]$, $t \in \{0, 1\}$:

$$Y_i(t) = \beta_0(t) + \beta_1(t)X_i^{(1)} + \cdots + \beta_d(t)X_i^{(d)} + \varepsilon_i(t) = \mathbf{x}_i^T\beta(t) + \varepsilon_i(t)$$

Where $\mathbf{x}_i = \begin{bmatrix} 1 \\ X_i^{(1)} \\ \vdots \\ X_i^{(d)} \end{bmatrix}$, $\beta(t) = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_d \end{bmatrix}$ and $\varepsilon_i(t) \in \mathbb{R}$ as the error term.

By stacking the n observations together we have:

$$\boldsymbol{Y}(t) = \boldsymbol{X}\beta(t) + \boldsymbol{\varepsilon}(t)$$

With $\boldsymbol{Y}(t) = \begin{bmatrix} Y_1(t) \\ \vdots \\ Y_n(t) \end{bmatrix}$, $\boldsymbol{X} = \begin{bmatrix} \mathbf{x}_1^T \\ \vdots \\ \mathbf{x}_n^T \end{bmatrix} = \begin{bmatrix} 1 & X_1^{(1)} & \dots & X_1^{(d)} \\ \vdots & \vdots & & \vdots \\ 1 & X_n^{(1)} & \dots & X_n^{(d)} \end{bmatrix}$, and $\boldsymbol{\varepsilon}(t) = \begin{bmatrix} \varepsilon_1(t) \\ \vdots \\ \varepsilon_n(t) \end{bmatrix}$

Under this linearity assumption, linear regression techniques can be employed to impute the missing potential outcomes, thereby enhancing estimation accuracy.

For concreteness, we will provide a walkthrough using the T-learner estimator. The process is closely similar for X-learners. For S-learners, under the consistency assumption, it suffices to denote the outcome variable as $Y = TY(0) + (1 - T)Y(1)$ and fit a single model on this expression, though this approach may lose linearity. For the T-Learner, an alternative expression of the ATE is given by:

$$\tau = \mathbb{E}[\mathbf{x}^T](\beta(1) - \beta(0))$$

This suggest an Ordinary Least Squares (OLS) estimator:

$$\hat{\tau}_{OLS} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i^T(\hat{\beta}(1) - \hat{\beta}(0))$$

Here, OLS gives us

$$\begin{aligned}\hat{\beta}(t) &= \underset{\beta \in \mathbb{R}^{d+1}}{\arg\min} \|\boldsymbol{Y}(t) - \boldsymbol{X}\beta\|_2 \\ &= \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{Y}(t)\end{aligned}$$

This provides an unbiased estimator for τ with reduced asymptotic variance compared to the Difference in Means estimator.

Furthermore, it can be demonstrated that even if the outcome does not exhibit a linear relationship with the covariates, the OLS estimator still possesses a smaller asymptotic variance compared to the DM estimator. [5]

4. Propensity Score based estimators

4.1 *The Propensity Score*

As we have mentionned before, in most practical scenarios, implementing a Randomized Controlled Trial can be challenging due to ethical considerations or practical limitations. When RCTs are not feasible, the ignorability assumption may be violated, as the likelihood of receiving treatment could differ based on the potential outcomes of the subjects. In such situations, we rely on the uncofoundedness assumption, which allows us to derive valid conclusions from observational data in the absence of randomization. Specifically, this section will address estimators that are based on the use of the *propensity score*.

The propensity score is defined as the probability that a given subject with covariates $X = x$ receives the active treatment, it can be expressed as a function e defined by:

$$\begin{aligned}e : \mathbb{R}^d &\longrightarrow [0,1] \\ x &\longmapsto \mathbb{P}[T = 1 \mid X = x]\end{aligned}$$

The main idea here is to utilize the propensity score to either group subpopulations with similar probabilities of receiving the treatment, or to use the calculated scores as weights to mitigate discrepancies.

Specifically, it can be demonstrated that conditioning on the propensity score of a subject establishes independence between the treatment and potential outcomes.

$$T \perp\!\!\!\perp \{Y(0), Y(1)\} \mid e(X)$$

In practice, the propensity score is not likely to be known and must be estimated from observed data. By assuming that $e(X)$ follows a parametric model, logic regression can be utilized to infer the propensity score. Formally, we assume that e is parameterized by $\beta = (\beta_0, \beta_1, \ldots, \beta_d) \in \mathbb{R}^{d+1}$ such that:

$$e(X, \beta) = \frac{1}{1 + e^{-\mathbf{x}^T\beta}}$$

Where, as in section 3.4, $\mathbf{x}^T = (1, X^{(1)}, \ldots, X^{(d)}) \in \mathbb{R}^{d+1}$.

The parameter β can be estimated using the maximum likelihood estimator, which consists of finding the maximum of the likelihood function:

$$\mathcal{L}(\beta) = \prod_{i=1}^{n} e(X_i, \beta)^{T_i} \left(1 - e(X_i, \beta)\right)^{1-T_i}$$

To achieve this, we determine the critical points of the log-likelihood function $\ell(\beta)$ by solving the following equation:

$$\frac{\partial \ell(\beta)}{\partial \beta_k} = \sum_{i=1}^{n} \frac{\partial e(X_i, \beta)}{\partial \beta_k} \left(\frac{T_i - e(X_i, \beta)}{e(X_i, \beta)\left(1 - e(X_i, \beta)\right)} \right) = 0 \text{ , for } k \in [\![0, d]\!]$$

The optimal parameter $\hat{\beta}$ is then obtained through numerical methods by solving for:

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^{d+1}}{\arg\max}\, \mathcal{L}(\beta)$$

Which in turn provides the estimated propensity score $e(X, \hat{\beta})$ for $\mathbb{P}[T = 1|X]$.

4.2 *Propensity Score Stratification*

Given $\hat{\beta}$, we obtain a set of statistics for the estimated propensity score for each subject: $(\hat{e}_i)_{i \in [\![1,n]\!]}$, where $\hat{e}_i = e(X_i, \hat{\beta})$ is the propensity score for the i-th subject.

From here, we divide the sample into strata based on the scores, aiming for each stratum to contain similar values of the covariates, thereby satisfying the conditional ignorability assumption. Formally, we first sort the observations in respect of their propensity scores:

$$\hat{e}_{(0)} \leqslant \ldots \leqslant \hat{e}_{(n)}$$

where $e_{(i)}$ represents the i-th smallest propensity score in the sample.

Utilizing the fact that the observations are exchangeable, we then define the boundaries for $K \ll n$ approximately equally sized strata using set of quantiles $(q_j)_{j \in [\![0,K]\!]}$, where $q_j \approx e_{(\lceil nj/K \rceil)}$.

Within each stratum j, we estimate the Average Treatment Effect using the Difference in Means estimator (c.f. 2.1). The final estimation $\hat{\tau}_{Strat}$ is obtained by a weighted sum of the ATEs of all the strata, where the weight is given by the proportion of observations within each strata. [3]

Defining $\mathcal{S}_j = \{i \in [\![1, n]\!] \mid \hat{e}_i \in]q_{j-1}, q_j]\}$, $K_j = |\mathcal{S}_j|$ and maintaining the notation of $\mathcal{T}_0$ and $\mathcal{T}_1$ in section 2.1, we denote $n_{j,0} = |\mathcal{S}_j \cap \mathcal{T}_0|$ and $n_{j,1} = |\mathcal{S}_j \cap \mathcal{T}_1|$. This gives us the following equality:

$$\hat{\tau}_{Strat} = \sum_{j=1}^{K} \frac{K_j}{K} \left(\frac{1}{n_{j,1}} \sum_{i \in \mathcal{S}_j \cap \mathcal{T}_1} Y_i - \frac{1}{n_{j,0}} \sum_{i \in \mathcal{S}_j \cap \mathcal{T}_0} Y_i \right)$$

4.3 *Inverse Propensity Score Weighting*

Instead of focusing on obtaining unbiased estimates within strata, weighting methods aim to provide an unbiased estimator for τ. The intuition is to use the propensity score to adjust for imbalances in the probability of treatment assignment, thereby approximating the conditions of a RCT. Using the consistency and unconfoundedness assumptions, we

derive the following: [2]

$$\begin{aligned}\mathbb{E}[Y(1)] &= \mathbb{E}_X\left[\mathbb{E}\left[Y(1) \mid X\right]\right] = \mathbb{E}_X\left[\frac{\mathbb{E}[T \mid X]\mathbb{E}\left[Y(1) \mid X\right]}{e(X)}\right] \\ &= \mathbb{E}_X\left[\frac{\mathbb{E}[TY(1) \mid X]}{e(X)}\right] = \mathbb{E}_X\left[\frac{\mathbb{E}[TY \mid X]}{e(X)}\right] \\ &= \mathbb{E}\left[\frac{TY}{e(X)}\right]\end{aligned}$$

In the second line, the unconfoundedness assumption is used to equate $\mathbb{E}[T \mid X]\mathbb{E}[Y \mid X]$ with $\mathbb{E}[TY \mid X]$. The consistency assumption is used along with the fact that $T(1-T) = 0$, resulting in $TY = T^2Y(1) + T(1-T)Y(0) = TY(1)$. Similarly for $\mathbb{E}[Y(0)]$, we obtain:

$$\mathbb{E}[Y(0)] = \mathbb{E}\left[\frac{(1-T)Y}{(1-e(X))}\right]$$

Thus, the Average Treatment Effect can be expressed as:

$$\tau = \mathbb{E}\left[\frac{TY}{e(X)} - \frac{(1-T)Y}{1-e(X)}\right]$$

The Inverse Propensity Score Weighting (IPW) estimator is therefore given by:

$$\hat{\tau}_{IPW} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{T_iY_i}{\hat{e}_i} - \frac{(1-T_i)Y_i}{1-\hat{e}_i}\right)$$

If the propensity score is known, then the Inverse Propensity Score Weighting (IPW) estimator is unbiased. Conversely, for the IPW estimator to be consistent, $e(X)$ must represent the true propensity score. The IPW estimator exhibits poor small-sample properties when the propensity score approaches 0 or 1. In such cases, units with very low or very high propensity scores can contribute extreme values to the estimate, leading to high variance and instability in the estimator.

4.4 *Augmented Inverse Propensity Score Weighting*

Due to the aforementioned limitations, the IPW estimator has been refined to integrate both the probability of treatment and predictive information about the outcome variable. This enhancement is realized through the Augmented Inverse Propensity Score Weighting (AIPW) estimator, which combines the IPW approach with a weighted average of the T-Learner to improve estimator performance. [2]

$$\hat{\tau}_{AIPW} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{T_iY_i}{\hat{e}_i} - \frac{T_i-\hat{e}_i}{\hat{e}_i}\hat{\mu}_1(X_i)\right) - \left(\frac{(1-T_i)Y_i}{1-\hat{e}_i} - \frac{T_i-\hat{e}_i}{1-\hat{e}_i}\hat{\mu}_0(X_i)\right)$$

As previously described, $\hat{\mu}_t$ are estimates from statistical models corresponding to the expectations of potentials outcomes $x \mapsto \mathbb{E}[Y \mid T = t, X = x]$ for $t \in \{0, 1\}$.

The AIPW estimator is referred to as "doubly robust" because it remains consistent

as long as either the treatment assignment mechanism or the outcome model is correctly specified. For instance, if the propensity score $\hat{e}$ accurately estimates the likelihood of a patient receiving the treatment of interest, then the term $T - \hat{e}$ will converge to zero in expectation, simplifying the AIPW estimator to the IPW estimator.

$$\hat{\tau}_{AIPW} = \frac{1}{n}\sum_{i=1}^{n} \underbrace{\left(\frac{T_i Y_i}{\hat{e}_i} - \frac{(1-T_i)Y_i}{1-\hat{e}_i}\right)}_{\text{IPW}} + \underbrace{\left(\left(1-\frac{T_i}{\hat{e}_i}\right)\hat{\mu}_1(X_i) - \left(1-\frac{1-T_i}{1-\hat{e}_i}\right)\hat{\mu}_0(X_i)\right)}_{\text{Augmentation}}$$

Conversely, if the conditional outcomes are correctly predicted, the estimator simplifies to the T-Learner.

$$\hat{\tau}_{AIPW} = \frac{1}{n}\sum_{i=1}^{n} \underbrace{(\hat{\mu}_1(X_i) - \hat{\mu}_0(X_i))}_{\text{T-Learner}} + \underbrace{\left(\frac{T_i}{\hat{e}_i}(Y_i - \hat{\mu}_1(X_i)) - \frac{1-T_i}{1-\hat{e}_i}(Y_i - \hat{\mu}_0(X_i))\right)}_{\text{Augmentation}}$$

References

[1] Sören R Künzel et al. "Metalearners for estimating heterogeneous treatment effects using machine learning". In: *Proceedings of the national academy of sciences* 116.10 (2019), pp. 4156–4165.

[2] Christoph F Kurz. "Augmented inverse probability weighting and the double robustness property". In: *Medical Decision Making* 42.2 (2022), pp. 156–167.

[3] Jared K Lunceford and Marie Davidian. "Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study". In: *Statistics in medicine* 23.19 (2004), pp. 2937–2960.

[4] Brady Neal. "Introduction to causal inference". In: *Course Lecture Notes (draft)* (2020).

[5] Stefan Wager. *Stats 361: Causal inference*. Tech. rep. Technical report, Technical report, Stanford University, 2020. URL: https ..., 2020.