



HAL
open science

Quantifying Treatment Effects: Estimating Risk Ratios in Causal Inference

Ahmed Boughdiri, Julie Josse, Erwan Scornet

► **To cite this version:**

Ahmed Boughdiri, Julie Josse, Erwan Scornet. Quantifying Treatment Effects: Estimating Risk Ratios in Causal Inference. 2024. hal-04588018

HAL Id: hal-04588018

<https://hal.science/hal-04588018>

Preprint submitted on 25 May 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Quantifying Treatment Effects: Estimating Risk Ratios in Causal Inference

Ahmed Boughdiri
Premedical project team
INRIA Sophia-Antipolis
Montpellier, France
ahmed.boughdiri@inria.fr

Julie Josse
Premedical project team
INRIA Sophia-Antipolis
Montpellier, France
julie.josse@inria.fr

Erwan Scornet
Sorbonne Université and Université Paris Cité, CNRS
Laboratoire de Probabilités, Statistique et Modélisation
F-75005 Paris, France
erwan.scornet@sorbonne-universite.fr

Abstract

Randomized Controlled Trials (RCT) are the current gold standards to empirically measure the effect of a new drug. However, they may be of limited size and resorting to complementary non-randomized data, referred to as observational, is promising, as additional sources of evidence. In both RCT and observational data, the Risk Difference (RD) is often used to characterize the effect of a drug. Additionally, medical guidelines recommend to also report the Risk Ratio (RR), which may provide a different comprehension of the effect of the same drug. While different methods have been proposed and studied to estimate the RD, few methods exist to estimate the RR. In this paper, we propose estimators of the RR both in RCT and observational data and provide both asymptotical and finite-sample analyses. We show that, even in an RCT, estimating treatment allocation probability or adjusting for covariates leads to lower asymptotic variance. In observational studies, we propose weighting and outcome modeling estimators and derive their asymptotic bias and variance for well-specified models. Using semi-parametric theory, we define two doubly robust estimators with minimal variances among unbiased estimators. We support our theoretical analysis with empirical evaluations and illustrate our findings through experiments.

1 Introduction

Motivation Modern evidence-based medicine prioritizes Randomized Controlled Trials (RCTs) as the cornerstone of clinical evidence. Randomization in RCTs allows for the estimation of the average treatment effect (ATE) by mitigating confounding influences from extraneous or undesirable factors. However, RCTs may face limitations due to stringent eligibility criteria, unrealistic real-world compliance, short study durations, and limited sample sizes. Medical journals such as JAMA (Bibbins-Domingo, 2024) and others (Hernan and Robins, 2016) have advocated the use of real-world data, often referred to as observational data, to provide additional sources of evidence. These data sets are typically less expensive to collect, more representative of the target population, and usually encompass large sample sizes. Various estimators have been proposed to measure the treatment effect on an absolute scale. In the context of RCTs, estimators such as the Horvitz and Thompson (1952) or the Difference-in-Means introduced by Splawa-Neyman et al. (1990) can be used to estimate the risk difference (RD). For observational studies, several methods aim to estimate the risk difference

while minimizing confounding effects. These methods include re-weighting using Inverse Probability Weighting (IPW) (Hirano et al., 2003), outcome modeling with the G-formula, and doubly robust approaches like Augmented Inverse Probability Weighting (AIPW) introduced by (Robins, 1986).

In scenarios where outcomes are always either positive or negative, relative measures such as the Risk Ratio (RR) become particularly relevant. As explained in the medical CONSORT (Moher et al., 2010) guidelines, it is crucial to report relative treatment effect measures like the Risk Ratio (RR) alongside absolute measures to provide a more comprehensive understanding of the effect and its implications, as neither measure alone offers a complete picture. Indeed selecting one measure over another carries several implications. Naylor et al. (1992) and Forrow et al. (1992) demonstrated that physicians' inclination to treat patients, based on their perception of therapeutic impact, is influenced by the scale utilized to present clinical effects. In addition, the treatment effect may be heterogeneous in one scale, i.e. the treatment effect varies according to patient characteristics, but homogeneous in another scale (Rothman, 2011), which significantly disrupts interpretation. Colnet et al. (2024b) discusses in depth causal measure's properties with a focus on generalization of the treatment effect from a trial to a target population. However, aside from Shirvaikar and Holmes (2023), who uses causal forests to estimate the RR for different subgroups in a population, no closed-form estimators currently exist to estimate the RR. This indicates that, in practice, highly heuristic techniques are used to report the Risk Ratio in RCTs. There is a clear gap and a need for robust estimators and comprehensive analyses of their properties to offer better assessments of treatment effects and follow medical recommendations even in observational studies.

Contributions In this paper, we propose and study various estimators of the Risk Ratio in both randomized controlled trials and observational studies. In Section 2, for RCTs, we introduce the Risk Ratio Neyman (RR-N) and the Risk Ratio Horvitz-Thompson (RR-HT) estimators. We derive the expressions of their biases and variances in both finite-sample and asymptotic settings. Our analysis gives the optimal exposure probability to minimize the variance and proves the benefits of estimating it even in an RCT setting. Regarding observational studies, we propose in Section 3 the Risk Ratio of Inverse Propensity Weighting (RR-IPW) and the Risk Ratio of outcome modeling (RR-G) and derive the expressions of their biases and variances in both finite-sample and asymptotic settings. We also establish central limit theorems. Based on semi-parametric theory (Kennedy, 2015), we establish two doubly robust estimators and their cross-fitted version: Risk Ratio Augmented Inverse Propensity Weighting (RR-AIPW) and Risk Ratio One-step (RR-OS). A direct consequence of semi-parametric theory is that RR-OS has the minimal variance among all unbiased estimators whereas RR-AIPW has also the same asymptotic variance but has a positive bias. Compared to the Risk Difference, studying the Risk Ratio poses additional technical difficulties, due to its non-linear nature. Dedicated mathematical tools are used based on linearization and semi-parametric theory. In Section 4, we substantiate our theoretical findings with empirical evaluations, demonstrating the practical relevance and robustness of our proposed estimators.

Problem setting In the potential outcome framework (see Rubin, 1974; Splawa-Neyman et al., 1990), we consider the random variables $(X, T, Y^{(0)}, Y^{(1)})$, where $X \in \mathbb{R}^p$ denotes covariates describing a patient, T is the treatment assignment ($T = 1$ when the treatment is given to an individual, $T = 0$ otherwise) and $Y^{(0)}$ (resp. $Y^{(1)}$) is the outcome of interest, describing the status of a patient without treatment (and with treatment respectively). In practice, we do not have access simultaneously to $Y^{(0)}$ and $Y^{(1)}$, and we only observe

$$Y = TY^{(1)} + (1 - T)Y^{(0)}.$$

Causal effect measures are functional of the joint distribution of potential outcomes (see Pearl, 2009). In particular, the Risk Difference (RD) and the Risk Ratio (RR) contrast the two states as followed:

$$\tau_{\text{RD}} = \mathbb{E}[Y^{(1)}] - \mathbb{E}[Y^{(0)}] \quad \text{and} \quad \tau_{\text{RR}} = \frac{\mathbb{E}[Y^{(1)}]}{\mathbb{E}[Y^{(0)}]}. \quad (1)$$

The aim of this paper is to propose and study estimators of τ_{RR} . To estimate this quantity, we assume to be given an i.i.d dataset $(X_1, T_1, Y_1), \dots, (X_n, T_n, Y_n)$.

2 Estimating the Risk Ratio in RCTs

Randomized Controlled Trials randomly assign treatment to patients in order to evaluate treatment effects. We focus on the Bernoulli design, one of the most widely used RCT designs (Rubin, 1974; Imbens and Rubin, 2015), where each participant has the same probability $e \in (0, 1)$ of being treated, independently of the treatments of others. Throughout the paper, we use the following assumptions.

Assumption 1 (Bernoulli Trial). *We consider the following assumptions:*

1. **Ignorability** or *Exchangeability*, that is: $T \perp\!\!\!\perp (Y^{(0)}, Y^{(1)})$.
2. **SUTVA** (*Stable Unit Treatment Value Assumption*): $Y = TY^{(1)} + (1 - T)Y^{(0)}$.
3. **i.i.d.** *The data set is i.i.d.* $(X_i, T_i, Y_i^{(0)}, Y_i^{(1)})_{i=1, \dots, n} \stackrel{i.i.d.}{\sim} \mathcal{P}$. *In particular, the treatment assignment of one participant does not influence that of another, that is, for all $i \neq j$, $T_i \perp\!\!\!\perp T_j$.*
4. **Trial positivity**: *Each participant i has a fixed probability $e \in (0, 1)$ of being assigned to the intervention group $\mathbb{P}[T_i = 1] = e$.*

Bernoulli trials offer the flexibility to easily include additional participants or handles dropouts, preserving the study's overall integrity. Besides, we need to guarantee the existence of our estimates.

Assumption 2 (Outcome positivity). *We suppose that both $Y^{(0)}$ and $Y^{(1)}$ are squared integrable and that $\mathbb{E}[Y^{(0)}|X], \mathbb{E}[Y^{(1)}|X] > 0$.*

We propose three RR estimators: the RR Horvitz-Thompson, the RR Neyman and the RR G-formula. We provide finite-sample as well as asymptotic bias and variance of these estimators.

2.1 Risk Ratio Horvitz-Thomson estimator (RR-HT)

The Horvitz-Thomson estimator (Horvitz and Thompson, 1952) for estimating the RD, applies inverse probability weighting to adjust for varying propensity scores (probability of being treated) across treated and control groups. We use the same strategy to define the RR Horvitz-Thomson estimator.

Definition 1 (Risk Ratio Horvitz-Thomson estimator). *Grant Assumption 1 and Assumption 2. The Risk Ratio Horvitz-Thomson estimator denoted $\hat{\tau}_{RR,HT,n}$ is defined as,*

$$\hat{\tau}_{RR,HT,n} = \frac{\sum_{i=1}^n \frac{T_i Y_i}{e}}{\sum_{i=1}^n \frac{(1-T_i) Y_i}{1-e}} \quad \text{if } \sum_{i=1}^n T_i < n \text{ and } 0 \text{ otherwise.} \quad (2)$$

Within the context of a Bernoulli trial, Proposition 1 proves that the Risk Ratio Horvitz-Thompson estimator is asymptotically unbiased and normally distributed.

Proposition 1 (Asymptotic normality of $\hat{\tau}_{RR,HT,n}$). *Under Assumption 1 and Assumption 2, the Risk Ratio Horvitz-Thompson estimator is asymptotically unbiased and satisfies*

$$\sqrt{n} (\hat{\tau}_{RR,HT,n} - \tau_{RR}) \xrightarrow{d} \mathcal{N}(0, V_{RR,HT}) \quad \text{where} \quad V_{RR,HT} = \tau_{RR}^2 \left(\frac{\mathbb{E}[(Y^{(1)})^2]}{e \mathbb{E}[Y^{(1)}]^2} + \frac{\mathbb{E}[(Y^{(0)})^2]}{(1-e) \mathbb{E}[Y^{(0)}]^2} \right). \quad (3)$$

If we assume that for all i , $M \geq Y_i \geq m > 0$ and $0 < \sum_{i=1}^n T_i < n$, we also have:

$$|\mathbb{E}[\hat{\tau}_{RR,HT,n}] - \tau_{RR}| \leq \frac{2M^3(1-e)^3}{nm^3e^3} \quad \text{and} \quad |\text{Var}(\hat{\tau}_{RR,HT,n}) - V_{RR,HT}| \leq \frac{4M^4(1-e)^6}{nm^6e^4}$$

Note that the asymptotic variance of the RR-HT is influenced by e , so one can minimize the variance of the estimator by choosing an optimal e_{opt} prior to conducting the RCT. e_{opt} is the solution of a polynomial equation which depends on whether $\text{Var}(Y^{(1)})$ or $\text{Var}(Y^{(0)})$ is larger. An estimation of the two variances thus lead to an estimation of the optimal e_{opt} , which equals 0.5 when both variance are equals (see Appendix A.1.1 for details).

2.2 Risk Ratio Neyman estimator (RR-N)

The Neyman estimator (Splawa-Neyman et al., 1990), also known as the difference-in-means estimator, calculates the difference between means of the outcome in treated and control groups. In order to estimate the RR, we compute the ratio of the means instead.

Definition 2 (Risk Ratio Neyman estimator). *Grant Assumption 1 and Assumption 2. Let $n_1 = \sum_{i=1}^n T_i$ and $n_0 = n - n_1$. The Risk Ratio Neyman estimator denoted $\hat{\tau}_{RR,N,n}$, is defined as*

$$\hat{\tau}_{RR,N,n} = \frac{\frac{1}{n_1} \sum_{i=1}^n T_i Y_i}{\frac{1}{n_0} \sum_{i=1}^n (1 - T_i) Y_i}, \quad \text{if the denominator is nonzero and 0 otherwise.} \quad (4)$$

As for the RD, the RR-N can be viewed as a variant of RR-HT, where the probability e to be treated (or propensity score) is estimated by $\hat{e} = n_1/n$.

Proposition 2 (Asymptotic normality of $\hat{\tau}_{RR,N,n}$). *Under Assumption 1 and Assumption 2, the Risk Ratio Neyman estimator is asymptotically unbiased and satisfies*

$$\sqrt{n} (\hat{\tau}_{RR,N,n} - \tau_{RR}) \xrightarrow{d} \mathcal{N}(0, V_{RR,N}) \quad \text{where} \quad V_{RR,N} = \tau_{RR}^2 \left(\frac{\text{Var}(Y^{(1)})}{e \mathbb{E}[Y^{(1)}]^2} + \frac{\text{Var}(Y^{(0)})}{(1-e) \mathbb{E}[Y^{(0)}]^2} \right) = V_{RR,HT} - \frac{\tau_{RR}^2}{e(1-e)}. \quad (5)$$

Furthermore, if we assume that for all i , $M \geq Y_i \geq m > 0$ and $0 < \sum_{i=1}^n T_i < n$, we have:

$$|\mathbb{E}[\hat{\tau}_{RR,HT,n}] - \tau_{RR}| \leq \frac{2M^3}{nm^3} \quad \text{and} \quad |\text{Var}(\hat{\tau}_{RR,HT,n}) - V_{RR,HT}| \leq \frac{4M^4}{nm^6}$$

The frequency of treatments assignments in the sample (for example, $\hat{e} = 0.48$) may be different from the actual probability of receiving treatment e . As Hirano et al. (2003); Hahn (1998); Robins et al. (1992) noticed for the RD, Proposition 2 demonstrates that for the estimation of the Risk Ratio, opting for \hat{e} over e (thereby employing the RR-N instead of the RR-HT) results in a reduced asymptotic variance, with a larger reduction when e is close to zero or one. This advantage arises because \hat{e} corresponds more closely to the actual probability of receiving treatment in the study. Similarly to the RR-HT, one can select an optimal value of e_{opt} at which the RCT should be conducted prior to the actual experiment to minimize the variance of the RR-N. This procedure is also in detailed in Appendix A.1.2.

2.3 Risk Ratio G-formula estimator (RR-G)

For all $x \in \mathbb{R}^p$, let $\mu_0^*(x) = \mathbb{E}[Y^{(0)}|X=x]$ and $\mu_1^*(x) = \mathbb{E}[Y^{(1)}|X=x]$ be the surface responses of the potential outcomes (Box and Wilson, 1992). Assume that we have at our disposal two estimators $\hat{\mu}_0(\cdot)$ and $\hat{\mu}_1(\cdot)$ which respectively estimate $\mu_0^*(\cdot)$ and $\mu_1^*(\cdot)$. We then employ the ratio of these two potential outcome estimations to compute the Risk Ratio. This method, termed the plug-in G-formula or outcome-based modeling, was first introduced by Robins (1986) for the Risk Difference.

Definition 3 (Ratio plug-in G-formula). *Given two estimators $\hat{\mu}_0(\cdot)$ and $\hat{\mu}_1(\cdot)$, the Risk Ratio G-formula estimator, denoted $\hat{\tau}_{RR,G}$, is defined as*

$$\hat{\tau}_{RR,G,n} = \frac{\sum_{i=1}^n \hat{\mu}_1(X_i)}{\sum_{i=1}^n \hat{\mu}_0(X_i)}, \quad \text{if } \sum_{i=1}^n \hat{\mu}_0(X_i) \neq 0 \text{ and zero otherwise.} \quad (6)$$

The properties of RR-G depend on the estimators $\hat{\mu}_0$ and $\hat{\mu}_1$. We analyze in the following the behavior of Oracle Risk Ratio G-formula estimator defined as $\tau_{RR,G}^* = (\sum_{i=1}^n \mu_1^*(X_i)) / (\sum_{i=1}^n \mu_0^*(X_i))$.

Proposition 3 (Asymptotic Normality of $\tau_{RR,G}^*$). *Grant Assumption 1 and Assumption 2. Then, the Oracle Risk Ratio G-formula estimator, $\tau_{RR,G}^*$, is asymptotically unbiased and satisfies*

$$\sqrt{n} (\tau_{RR,G,n}^* - \tau_{RR}) \xrightarrow{d} \mathcal{N}(0, V_{RR,G}) \quad \text{and} \quad V_{RR,G} = \tau_{RR}^2 \text{Var} \left(\frac{\mu_1^*(X)}{\mathbb{E}[Y^{(1)}]} - \frac{\mu_0^*(X)}{\mathbb{E}[Y^{(0)}]} \right) \quad (7)$$

Assume for all i , $M_0 \geq \mu_0^*(x_i) \geq m_0 > 0$ and $M_1 \geq \mu_1^*(x_i)$. Then, we have:

$$|\mathbb{E}[\hat{\tau}_{RR,G,n}^*] - \tau_{RR}| \leq \frac{2M_1 M_0^2}{nm_0^3} \quad \text{and} \quad |\text{Var}(\hat{\tau}_{RR,G,n}^*) - V_{RR,G}| \leq \frac{2M_0^2 M_1 (M_1 + M_0)}{nm_0^6}$$

Proposition 3 establishes that the Oracle Risk Ratio G-formula estimator is asymptotically normal. Surprisingly, in the case where there is no effect (i.e. $\tau_{RR} = 1$), the asymptotic variance is driven by the variance of the Risk Difference on each strata determined by X , namely $\text{Var}(\mu_1^*(X) - \mu_0^*(X))$. By considering the Oracle RR-G instead of RR-G, we remove the additional randomness related to the estimation of the surface responses. It is thus likely that the true variance of RR-G is larger than that of Oracle RR-G.

In practice, comparing the performances of $\hat{\tau}_{RR,N}$ and $\hat{\tau}_{RR,G}$ is difficult as (i) we do not have the asymptotic variance of $\hat{\tau}_{RR,G}$ with estimated response surfaces functions but only that of the oracle RR-G and (ii) the resulting variance of the RR-G does not involve e (whose role is hidden in the estimators $\hat{\mu}_0$ and $\hat{\mu}_1$). However, if we assume a functional form for the surface responses as a linear one, one can show that even in an RCT setting, adjusting for covariates is beneficial as the variance of the oracle RR-G is smaller than the variance of RR-N. The exact result is provided in Appendix A.1.4.

3 Observational study estimators of the Risk Ratio

Observational studies reveal the complexities of real-world scenarios, which may be missed by the controlled designs of RCTs. A key difference between RCTs and observational studies is the handling of confounding variables. If not properly addressed, these can distort the true association between exposure and outcome due to their correlation with both. Therefore, estimating the Risk Ratio in observational studies is more complex than in RCTs, as randomization assumptions do not apply (i.e. the propensity score now depends on the covariates X).

Assumption 3 (Observational study identifiability assumptions). *We have*

1. **Unconfoundedness or Conditional Exchangeability:** $(Y(0), Y(1)) \perp\!\!\!\perp T \mid X$.
2. **Overlap or Positivity,** $\exists \eta \in (0, 1/2)$ such that, almost surely, $\eta \leq \mathbb{P}[T = 1 \mid X] \leq 1 - \eta$.
3. **SUTVA (Stable Unit Treatment Value Assumption)** holds: $Y = TY^{(1)} + (1 - T)Y^{(0)}$.
4. **i.i.d.** We still assume that the data set is i.i.d. $(X_i, T_i, Y_i^{(0)}, Y_i^{(1)})_{i=1, \dots, n} \stackrel{i.i.d.}{\sim} \mathcal{P}$.

Unconfoundedness means that after accounting for known confounding variables, no hidden factors affect both treatment assignment and outcomes. It is a relaxed form of exchangeability. Overlap ensures that the propensity score, $e(X) = \mathbb{P}[T = 1 \mid X]$ is never 0 or 1. We introduce three estimators for estimating the RR in observational studies. The Risk Ratio G-formula introduced in Section 2 can be used in the context of observational studies, contrary to the RR Horvitz-Thomson or RR Neyman estimators, which are built on the assumption of a constant propensity score. However, the RR Neyman estimator can be extended to observational studies as follows.

3.1 Risk Ratio Inverse Propensity Weighting (RR-IPW)

The Inverse Propensity Weighting (IPW) (Hirano et al., 2003), for estimating the RD, can be interpreted as a simple difference-in-means, where individuals are weighted by the inverse of their propensity score. The same idea can be used to build the RR-IPW. However, despite the similarities between the approaches, proofs are very different between RD and RR, and much more involved with the ratio. This remarks applies to all estimators presented.

Definition 4 (RR-IPW). *Grant Assumption 2 and Assumption 3. Given an estimator $0 < \hat{e}(\cdot) < 1$ of the propensity score $e(x) = \mathbb{P}[T = 1 \mid X = x]$, the Risk Ratio IPW, denoted by $\hat{\tau}_{RR,IPW}$, is defined as*

$$\hat{\tau}_{RR,IPW} = \frac{\sum_{i=1}^n \frac{T_i Y_i}{\hat{e}(X_i)}}{\sum_{i=1}^n \frac{(1-T_i) Y_i}{1-\hat{e}(X_i)}}.$$

Proposition 4 demonstrates the asymptotic normality of the Oracle Ratio IPW estimator, defined as the RR-IPW but where $\hat{e}(\cdot)$ is replaced by the oracle propensity score $e(\cdot)$.

Proposition 4 (RR-IPW asymptotic normality). *Grant Assumption 2 and Assumption 3. Then the Oracle Risk Ratio IPW defined above is asymptotically unbiased and satisfies*

$$\sqrt{n} (\tau_{RR,IPW}^* - \tau_{RR}) \xrightarrow{d} \mathcal{N}(0, V_{RR,IPW}) \quad \text{and} \quad V_{RR,IPW} = \tau_{RR}^2 \left(\frac{\mathbb{E} \left[\frac{(Y^{(1)})^2}{e(X)} \right]}{\mathbb{E} [Y^{(1)}]^2} + \frac{\mathbb{E} \left[\frac{(Y^{(0)})^2}{1-e(X)} \right]}{\mathbb{E} [Y^{(0)}]^2} \right). \quad (8)$$

Furthermore, if for all i , $0 < m \leq Y_i \leq M$ we have:

$$|\mathbb{E}[\tau_{RR,IPW}^*] - \tau_{RR}| \leq \frac{2M^3(1-\eta)^3}{nm^3\eta^3} \quad \text{and} \quad |\text{Var}(\tau_{RR,IPW}^*) - V_{RR,IPW}| \leq \frac{4M^4(1-\eta)^6}{nm^6\eta^4}, \quad (9)$$

Note that when the propensity score is constant, one can retrieve the variance of the RR-HT as expected. Note also that the asymptotic variance may be large, due to strata on which the propensity score is close to zero or one. In other words, a correct estimation is difficult when some subpopulations are unlikely to be treated (or untreated).

3.2 Risk Ratio Augmented Inverse Propensity Weighting (RR-AIPW)

A popular estimator for the RD is the augmented inverse probability weighted estimator (AIPW). It combines both the properties of the regression based estimator (G-formula) and the inverse probability weighted estimator (IPW) and is 'doubly robust' in the sense that it only requires either the propensity or outcome model to be correctly specified. We can directly use the AIPW methodology for the numerator and denominator of the Risk Ratio separately, which when combined with cross-fitting (Chernozhukov et al., 2017), leads to the following definition.

Definition 5 (Risk Ratio AIPW crossfitted). We denote $\mathcal{I} = \{1, \dots, n\}$, let $\mathcal{I}_1, \mathcal{I}_2, \dots, \mathcal{I}_K$ be a partition of \mathcal{I} . We assumed to be given an estimator $0 < \hat{e}(\cdot) < 1$ of the propensity score $e(x) = \mathbb{P}[T = 1|X = x]$ and two estimators $\hat{\mu}_0(\cdot)$ and $\hat{\mu}_1(\cdot)$ of the surface responses $\mu_0(x) = \mathbb{E}[Y^{(0)}|X = x]$ and $\mu_1(x) = \mathbb{E}[Y^{(1)}|X = x]$. Then, the Risk Ratio AIPW crossfitted is defined as

$$\tilde{\tau}_{RR,AIPW} = \frac{\sum_{k=1}^K \sum_{i \in \mathcal{I}_k} \left(\hat{\mu}_{(1)}^{\mathcal{I}-k}(X_i) + \frac{T_i \cdot (Y_i - \hat{\mu}_{(1)}^{\mathcal{I}-k}(X_i))}{\hat{e}^{\mathcal{I}-k}(X_i)} \right)}{\sum_{k=1}^K \sum_{i \in \mathcal{I}_k} \left(\hat{\mu}_{(0)}^{\mathcal{I}-k}(X_i) + \frac{(1-T_i) \cdot (Y_i - \hat{\mu}_{(0)}^{\mathcal{I}-k}(X_i))}{1 - \hat{e}^{\mathcal{I}-k}(X_i)} \right)} := \frac{\tilde{\tau}_{RR,AIPW,1}}{\tilde{\tau}_{RR,AIPW,0}}$$

where $\hat{\mu}_{(t)}^{\mathcal{I}-k}(X)$ and $\hat{e}^{\mathcal{I}-k}(X)$ are the estimates of $\mu_{(t)}$ and e built on the sample $\mathcal{I}_{-k} = \mathcal{I} \setminus \mathcal{I}_k$.

Proposition 5 (Risk Ratio AIPW asymptotic normality). Grant Assumption 2 and Assumption 3. Then, if we assume that for all $1 \leq k \leq K$ and all $t \in \{0, 1\}$,

$$\mathbb{E} \left[\left(\hat{\mu}_{(t)}^{\mathcal{I}-k}(X_i) - \mu_{(t)}(X_i) \right)^2 \right] \mathbb{E} \left[\left(\hat{e}^{\mathcal{I}-k}(X_i) - e(X_i) \right)^2 \right] = o \left(\frac{1}{n} \right) \quad (10)$$

$$\mathbb{E} \left[\left(\hat{\mu}_{(t)}^{\mathcal{I}-k}(X_i) - \mu_{(t)}(X_i) \right)^2 \right] = o_{\mathbb{P}}(1) \quad \text{and} \quad \mathbb{E} \left[\left(\hat{e}^{\mathcal{I}-k}(X) - e(X) \right)^2 \right] = o_{\mathbb{P}}(1), \quad (11)$$

we have that the crossfitted Risk Ratio AIPW estimator is asymptotically unbiased and satisfies

$$\sqrt{n} (\tilde{\tau}_{RR,AIPW} - \tau_{RR}) \xrightarrow{d} \mathcal{N}(0, V_{RR,AIPW})$$

where

$$\frac{V_{RR,AIPW}}{\tau_{RR}^2} = \text{Var} \left(\frac{\mu_1(X)}{\mathbb{E}[Y^{(1)}]} - \frac{\mu_0(X)}{\mathbb{E}[Y^{(0)}]} \right) + \mathbb{E} \left[\frac{\text{Var}(Y^{(1)}|X)}{e(X)\mathbb{E}[Y^{(1)}]^2} \right] + \mathbb{E} \left[\frac{\text{Var}(Y^{(0)}|X)}{(1-e(X))\mathbb{E}[Y^{(0)}]^2} \right]. \quad (12)$$

Note that we employ the same assumptions as those used to study the Risk Difference AIPW (Wager, 2020), which is crucial to achieve double robustness. Additionally, Assumption 10, also known as Risk decay, is verified when one estimation of the nuisance component achieves a parametric rate, while the other estimation is consistent.

3.3 Risk Ratio One-step estimator (RR-OS)

Augmented Inverse Probability Weighting (AIPW) estimator is derived from semi-parametric theory (Tsiatis, 2006), in particular by calculating the influence function for the statistical estimand $\psi = \mathbb{E}[\mathbb{E}[Y|T = 1, X] - \mathbb{E}[Y|T = 0, X]]$ for the Risk Difference and constructing a one-step estimator (Kennedy, 2015). Such estimator exhibits the lowest mean squared error among all estimators and

therefore the smallest variance among all unbiased estimators. The RR-AIPW corresponds to a ratio of two one-step estimators, one for $\mathbb{E}[Y^{(1)}]$ and one for $\mathbb{E}[Y^{(0)}]$. By applying the influence function theory to the Risk Ratio itself (and not to the numerator and denominator separately), we obtain the following estimator.

Definition 6. *The crossfitted One-step estimator $\hat{\tau}_\varphi$ is defined as*

$$\hat{\tau}_\varphi = \frac{\sum_{k=1}^K \sum_{i \in \mathcal{I}_k} \hat{\mu}_{(1)}^{\mathcal{I}-k}(X_i)}{\sum_{k=1}^K \sum_{i \in \mathcal{I}_k} \hat{\mu}_{(0)}^{\mathcal{I}-k}(X_i)} \left(1 - \frac{\tilde{\tau}_{RR,AIPW,1}}{\sum_{k=1}^K \sum_{i \in \mathcal{I}_k} \hat{\mu}_{(0)}^{\mathcal{I}-k}(X_i)} \right) + \frac{\tilde{\tau}_{RR,AIPW,0}}{\sum_{k=1}^K \sum_{i \in \mathcal{I}_k} \hat{\mu}_{(0)}^{\mathcal{I}-k}(X_i)}, \quad (13)$$

where $\tilde{\tau}_{RR,AIPW,1}$ and $\tilde{\tau}_{RR,AIPW,0}$ are defined in 5.

Proposition 6 (Asymptotic normality/ variance). *Grant Assumption 2, Assumption 3 and assume that Equation (10) holds. Then, if we assume that for all $1 \leq k \leq K$,*

$$\mathbb{E} \left[\hat{\mu}_0^{\mathcal{I}-k}(X) \right] - \mathbb{E} \left[\mu_0^{\mathcal{I}-k}(X) \right] = o_p \left(n^{-1/4} \right) \quad (14)$$

$$\mathbb{E} \left[\left(\hat{\mu}_0^{\mathcal{I}-k}(X) - \mu_0^{\mathcal{I}-k}(X) \right)^2 \right] \mathbb{E} \left[\left(\hat{\mu}_1^{\mathcal{I}-k}(X) - \mu_1^{\mathcal{I}-k}(X) \right)^2 \right] = o_p \left(n^{-1} \right), \quad (15)$$

the One-Step estimator is asymptotically unbiased and satisfies

$$\sqrt{n} (\hat{\tau}_\varphi - \tau_{RR}) \xrightarrow{d} \mathcal{N} (0, V_{RR,AIPW}),$$

where $V_{RR,AIPW}$ is defined in Proposition 5.

It is notable that RR-OS and RR-AIPW have the same variance. However, despite RR-OS being a more complex estimator that requires additional assumptions, it has the advantage of having no first-order finite sample bias. Moreover, RR-OS actually achieves the minimal variance among all unbiased estimators of the RR (see Tsiatis, 2006).

4 Simulation

Simulations for randomized controlled trials are provided in Appendix B. For observational studies, we generate datasets $(X_i, T_i, Y_i^{(0)}, Y_i^{(1)})$ according to the following model:

$$\begin{aligned} Y_i^{(1)} &= m(X_i) + b(X_i) + \varepsilon_i^{(1)} \\ Y_i^{(0)} &= b(X_i) + \varepsilon_i^{(0)} \end{aligned} \quad \mathbb{P}[T_i = 1 | X_i] = e(X_i),$$

with $\varepsilon^{(t)} \sim \mathcal{N}(0, \sigma^2)$. Each of the following setups have specificities regarding the $m(\cdot)$, $b(\cdot)$, and $e(\cdot)$ functions, which respectively correspond to the treatment effect, the baseline and propensity score. The true Risk Ratio can be expressed as $\tau_{RR} = \mathbb{E}[Y^{(1)}] / \mathbb{E}[Y^{(0)}] = \mathbb{E}[m(X)] / \mathbb{E}[b(X)] + 1$. We compare the performances of all estimators defined in Section 3 where nuisance components (regression surfaces and propensity score) are estimated via parametric (linear/logistic regression) or non parametric methods (random forests). Each simulation is repeated 3000 times. Further settings and details are provided in appendix B.

4.1 Linear and Logistic DGP

The first observational data generating process (DGP) is a parametric setup introduced in Lunceford and Davidian (2004), composed of linear outcome models (linear treatment effect and baseline) and a logistic propensity score, that is

$$\begin{aligned} m(X, V) &= 2, \\ b(X, V) &= \beta_0^\top [X, V], \quad \text{where } \beta_0 = (-1, 1, -1, -1, 1, 1) \\ e(X) &= (1 + \exp(-\beta_e X))^{-1}, \quad \text{where } \beta_e = (-0.6, 0.6, -0.6). \end{aligned}$$

The covariates $\mathbf{X} = (X_1, X_2, X_3)^\top$ are associated with both treatment exposure and potential response while $\mathbf{V} = (V_1, V_2, V_3)^\top$ are associated with the potential response but not directly related to treatment exposure. $[X, V]$ follow a joint distribution by taking $X_3 \sim \text{Bernoulli}(0.2)$ and

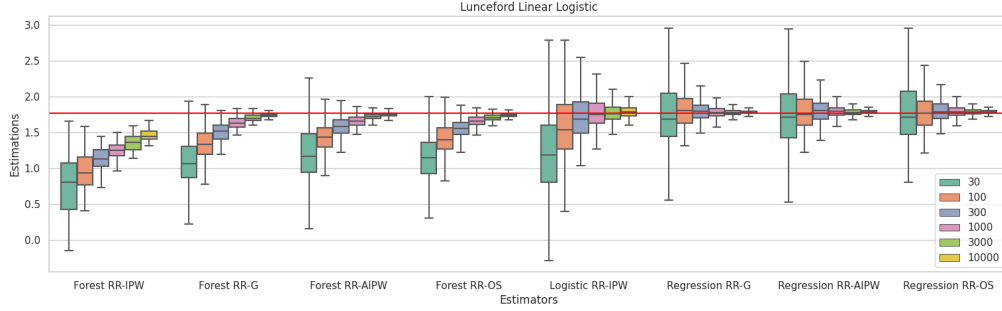


Figure 1: Estimations of the Risk Ratio with weighting, outcome based and augmented estimators as a function of the sample size for the Linear-Logistic DGP. Parametric (Regression) and non parametric (Forest) estimations of nuisance are displayed.

then generating V_3 as Bernoulli with $P(V_3 = 1 | X_3) = 0.75X_3 + 0.25(1 - X_3)$. Conditional on X_3 , $(X_1, V_1, X_2, V_2)^T$ was then generated as multivariate normal $\mathcal{N}(\lambda_{X_3}, \Sigma_{X_3})$, where

$$\lambda_1 = \begin{pmatrix} 1 \\ 1 \\ -1 \\ -1 \end{pmatrix} \lambda_0 = \begin{pmatrix} -1 \\ -1 \\ 1 \\ 1 \end{pmatrix} \quad \text{and} \quad \Sigma_1 = \Sigma_0 = \begin{pmatrix} 1 & 0.5 & -0.5 & -0.5 \\ 0.5 & 1 & -0.5 & -0.5 \\ -0.5 & -0.5 & 1 & 0.5 \\ -0.5 & -0.5 & 0.5 & 1 \end{pmatrix}.$$

Results are depicted in Figure 1. Only confounding variables are used. As expected since the generative process is linear, methods that use parametric estimators (logistic/linear regression) outperform those using non-parametric approaches (random forests) in finite-sample settings. While all methods (except maybe Forest RR-IPW) converges to the correct RR, methods based on parametric estimators exhibit a faster rate of convergence and are unbiased (except for Logistic RR-IPW) even for small sample sizes. Indeed, random forests are not suited for linear generative process and require here more than 10000 samples to estimate correctly the RR. All in all, when the outcome modelling and the propensity scores are linear, the two doubly-robust estimators (RR-AIPW and RR-OS) and the RR-G, all based on linear estimators, achieve the best performances: they are unbiased, even for small sample sizes, and converge quickly to the true RR. Furthermore, both Regression RR-OS and RR-AIPW estimators give very similar results.

4.2 Non-Linear and Non-Logistic DGP

We use the same simulations as in Nie and Wager (2020) using nonlinear models for every quantity, as detailed below

$$\begin{aligned} m(X) &= \sin(\pi X_1 X_2) + 2(X_3 - 0.5)^2 + X_4 + 0.5X_5 - (X_1 + X_2)/4 \\ b(X) &= (X_1 + X_2)/2 \\ e(X) &= \max\{0.1, \min(\sin(\pi X_1), 0.9)\}, \quad \text{where } X \sim \text{Unif}(0, 1)^6. \end{aligned}$$

Results are presented in Figure 2. At first glance, all methods seem to have similar performances. However, estimators based on parametric estimators (last four) fail to converge to the correct quantity. They present an intrinsic bias, which does not vanish as the sample size increases. This was expected as linear methods are unable to model the complex non-linear generative process of this simulation. On the other hand, methods that employ random forests estimators achieve good performances: they are consistent and unbiased even for small sample sizes. Note that RIPW has a larger variance than the other methods, with a small bias for very small sample sizes. Therefore, the G-formula and the two doubly-robust estimators that use random forests are competitive in this setting. Here again, both double robust estimators give similar performances.

4.3 Non-Linear and Logistic DGP

We use here a semi-parametric setup introduced by Nie and Wager (2020) with non-linear baseline models, a constant treatment effect and a logistic propensity score:

$$b(X) = 2 \log(1 + e^{X_{i1} + X_{i2} + X_{i3}}), \quad m(X) = 1, \quad \text{and} \quad e(X) = 1 / (1 + e^{X_{i2} + X_{i3}}),$$

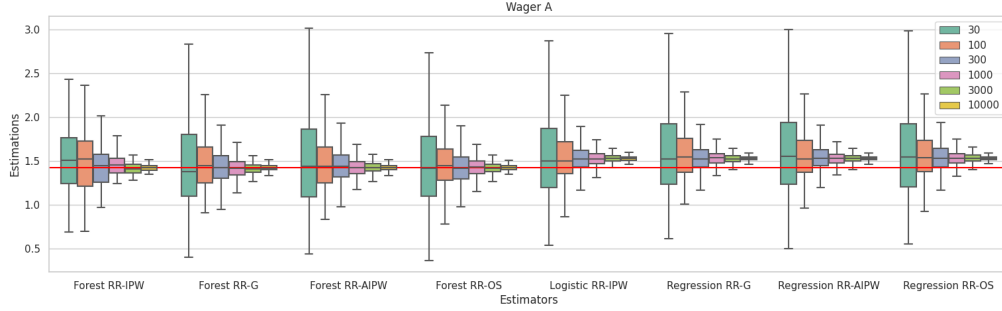


Figure 2: Estimations of the Risk Ratio with weighting, outcome based and augmented estimators as a function of the sample size for the non-Linear-non-Logistic DGP. Parametric (Regression) and non parametric (Forest) estimations of nuisance are displayed.

where $X_i \sim \mathcal{N}(0, I_{d \times d})$. Results are presented in Figure 3. Two estimators yield poor RR estimates

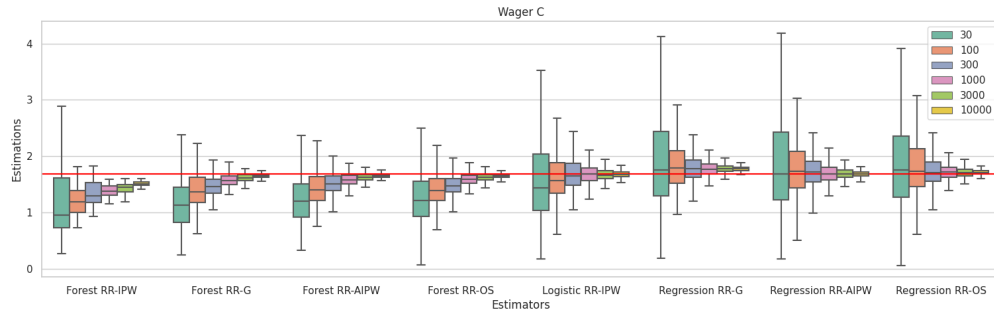


Figure 3: Estimations of the Risk Ratio with weighting, outcome based and augmented estimators as a function of the sample size for the non-Linear-Logistic DGP. Parametric (Regression) and non parametric (Forest) estimations of nuisance are displayed.

for the largest sample size: the forest IPW and the Regression G-formula. The forest IPW uses a random forest to estimate a logistic model, which may still converge but at a slower rate compared to other methods. The Regression G-formula employs regressions to approximate the response surfaces, potentially leading to an irreducible asymptotic bias. The Random Forest G-formula, Forest AIPW, and Forest OS estimators converge slowly to the true RR. This simulation highlights the double robust properties of the Regression RAIW and Regression ROS estimators, which target the true RR even at small sample sizes, as they have at least one well-specified model.

5 Conclusion

Quantifying treatment effects presents challenges, since different measures may lead to different understanding of the same phenomenon. In our study, we focus on one of these measures, the Risk Ratio and introduced several estimators, valid in RCT or observational studies. Using dedicated mathematical tools (linearization, semi-parametric theory), we establish their theoretical properties (controlled biases, consistency, central limit theorems). Empirical evaluations show that in complete linear (or nonlinear) settings, the estimating method has less importance than the estimation procedure used for the nuisance components (surface responses or propensity scores). In a more nuanced scenario, nonparametric estimation of nuisance components leads to larger variance and smaller biases compared to linear estimations. Empirical analyses have shown no difference between the two double-robust estimators, so we might recommend using the simplest one cross-fitted RR AIPW as it requires less assumptions.

Identifying guidelines establishing when linear nuisance components should be used instead of nonparametric ones still remains an open problem. Besides, while our theoretical analysis sheds some lights on estimator performances, we were only able to study the behavior of the oracle G-formula.

The additional source of randomness, resulting from the estimation of the nuisance components, should be taken into account in future works. In practice, observational studies may be used to generalize the treatment effect from a RCT population to the general population of interest. Our work is a first step toward proposing procedures to generalize the Risk Ratio to general populations.

References

- Bibbins-Domingo, K. (2024, 05). Causal inference about the effects of interventions from observational studies in medical journals. *JAMA*.
- Box, G. E. P. and K. B. Wilson (1992). *On the Experimental Attainment of Optimum Conditions*, pp. 270–310. New York, NY: Springer New York.
- Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins (2017). Double/debiased machine learning for treatment and causal parameters.
- Colnet, B., J. Josse, G. Varoquaux, and E. Scornet (2024a). Reweighting the rct for generalization: finite sample error and variable selection.
- Colnet, B., J. Josse, G. Varoquaux, and E. Scornet (2024b). Risk ratio, odds ratio, risk difference... which causal measure is easier to generalize? *arXiv preprint arXiv:2303.16008*.
- Derumigny, A., L. Girard, and Y. Guyonvarch (2019). On the construction of confidence intervals for ratios of expectations.
- Forrow, L., W. C. Taylor, and R. M. Arnold (1992). Absolutely relative: How research results are summarized can affect treatment decisions. *The American Journal of Medicine* 92(2), 121–124.
- Hahn, J. (1998). On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica* 66(2), 315–332.
- Hernan, M. and J. Robins (2016, 03). Using big data to emulate a target trial when a randomized trial is not available: Table 1. *American Journal of Epidemiology* 183, kwv254.
- Hirano, K., G. Imbens, and G. Ridder (2003, 02). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica* 71, 1161–1189.
- Horvitz, D. G. and D. J. Thompson (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association* 47(260), 663–685.
- Imbens, G. W. and D. B. Rubin (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press.
- Kennedy, E. (2015, 10). Semiparametric theory and empirical processes in causal inference.
- Lunceford, J. and M. Davidian (2004, 10). Stratification and weighting via the propensity score in estimation of causal treatment effects: A comparative study. *Statistics in medicine* 23, 2937–60.
- Moher, D., S. Hopewell, K. F. Schulz, V. Montori, P. C. Gøtzsche, P. J. Devereaux, D. Elbourne, M. Egger, and D. G. Altman (2010). Consort 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials. *BMJ* 340.
- Naylor, C. D., E. Chen, and B. Strauss (1992, December). Measured enthusiasm: does the method of reporting trial results alter perceptions of therapeutic effectiveness? *Ann. Intern. Med.* 117(11), 916–921.
- Nie, X. and S. Wager (2020). Quasi-oracle estimation of heterogeneous treatment effects.
- Pearl, J. (2009). *Causality* (2 ed.). Cambridge University Press.
- Robins, J. (1986). A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Mathematical Modelling* 7(9), 1393–1512.

- Robins, J. M., S. D. Mark, and W. Newey (1992). Estimating exposure effects by modelling the expectation of exposure conditional on confounders. *Biometrics* 48 2, 479–95.
- Rothman, K. J. (2011). *Epidemiology: an introduction* (2 ed.). Oxford University Press.
- Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational psychology* 66(5), 688–701.
- Shirvaikar, V. and C. Holmes (2023). Targeting relative risk heterogeneity with causal forests.
- Splawa-Neyman, J., D. M. Dabrowska, and T. P. Speed (1990). On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9. *Statistical Science* 5(4), 465 – 472.
- Stefanski, L. A. and D. D. Boos (2002). The calculus of m-estimation. *The American Statistician* 56(1), 29–38.
- Tsiatis, A. (2006, 01). Semiparametric theory and missing data. 73.
- Vaart, A. W. v. d. (1998). *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.
- Wager, S. (2020). Stats 361: Causal inference. Technical report, Technical report, Stanford University, 2020. URL: <https://web.stanford.edu> . . .

A Appendix / supplemental material

Since we are studying the asymptotic properties of the risk ratio, we cannot directly apply a central limit theorem as in Wager (2020). We will therefore rely on Theorem 1 to prove most of our asymptotic results.

Theorem 1 (Asymptotic normality of the ratio of two estimators). *Let (Z_1, \dots, Z_n) be n i.i.d. random variables, g_0 and g_1 two functions square integrable such that $\mathbb{E}[g_0(Z_i)] = \tau_0$ and $\mathbb{E}[g_1(Z_i)] = \tau_1$, where $\tau_0 \neq 0$. Then, we have that*

$$\sqrt{n} \left(\frac{\sum_{i=1}^n g_1(Z_i)}{\sum_{i=1}^n g_0(Z_i)} - \frac{\tau_1}{\tau_0} \right) \xrightarrow{d} \mathcal{N}(0, V_{RR}^*),$$

where

$$V_{RR}^* = \left(\frac{\tau_1}{\tau_0} \right)^2 \text{Var} \left(\frac{g_1(Z)}{\tau_1} - \frac{g_0(Z)}{\tau_0} \right).$$

Proof. We rely on M-estimation theory to prove Theorem 1. Let

$$\hat{\theta}_n = \begin{pmatrix} \frac{1}{n} \sum_{i=1}^n g_0(Z_i) \\ \frac{1}{n} \sum_{i=1}^n g_1(Z_i) \\ \frac{\sum_{i=1}^n g_1(Z_i)}{\sum_{i=1}^n g_0(Z_i)} \end{pmatrix} \quad \text{and} \quad \psi(Z, \theta) = \begin{pmatrix} g_0(Z) - \theta_0 \\ g_1(Z) - \theta_1 \\ \theta_1 - \theta_2 \theta_0 \end{pmatrix}, \quad (16)$$

where $\theta = (\theta_0, \theta_1, \theta_2)$. We have that

$$\sum_{i=1}^n \left(g_0(Z_i) - \frac{1}{n} \sum_{j=1}^n g_0(Z_j) \right) = \sum_{i=1}^n g_0(Z_i) - \sum_{j=1}^n g_0(Z_j) = 0,$$

and similarly

$$\sum_{i=1}^n \left(g_1(Z_i) - \frac{1}{n} \sum_{j=1}^n g_1(Z_j) \right) = \sum_{i=1}^n g_1(Z_i) - \sum_{j=1}^n g_1(Z_j) = 0.$$

Besides,

$$\sum_{i=1}^n \left(\frac{1}{n} \sum_{j=1}^n g_1(Z_j) - \frac{\sum_{j=1}^n g_1(Z_j)}{\sum_{j=1}^n g_0(Z_j)} \frac{1}{n} \sum_{j=1}^n g_0(Z_j) \right) = 0.$$

Gathering the three previous equalities, we obtain

$$\sum_{i=1}^n \psi(Z_i, \boldsymbol{\theta}_n) = 0, \quad (17)$$

which proves that $\hat{\boldsymbol{\theta}}_n$ is an M-estimator of type ψ (see Stefanski and Boos, 2002). Furthermore, letting $\boldsymbol{\theta}_\infty = (\tau_0, \tau_1, \tau_1/\tau_0)$, simple calculations show that

$$\mathbb{E}[\psi(Z, \boldsymbol{\theta}_\infty)] = 0. \quad (18)$$

Since the first two components of ψ are linear with respect to θ_0 and θ_1 and since the third component is linear with respect to θ_2 , $\boldsymbol{\theta}_\infty$ defined above is the only value satisfying (18). Define

$$A(\boldsymbol{\theta}_\infty) = \mathbb{E} \left[\frac{\partial \psi}{\partial \boldsymbol{\theta}} \Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_\infty} \right] \quad \text{and} \quad B(\boldsymbol{\theta}_\infty) = \mathbb{E} [\psi(Z, \boldsymbol{\theta}_\infty) \psi(Z, \boldsymbol{\theta}_\infty)^T]. \quad (19)$$

We now check the conditions of Theorem 7.2 in Stefanski and Boos (2002). First, let us compute $A(\boldsymbol{\theta}_\infty)$ and $B(\boldsymbol{\theta}_\infty)$. Since

$$\frac{\partial \psi}{\partial \boldsymbol{\theta}}(Z, \boldsymbol{\theta}) = \begin{pmatrix} -1 & 0 & 0 \\ 0 & -1 & 0 \\ -\theta_2 & 1 & -\theta_0 \end{pmatrix}, \quad (20)$$

we obtain

$$A(\boldsymbol{\theta}_\infty) = \begin{pmatrix} -1 & 0 & 0 \\ 0 & -1 & 0 \\ -\frac{\tau_1}{\tau_0} & 1 & -\tau_0 \end{pmatrix}, \quad (21)$$

which leads to

$$A^{-1}(\boldsymbol{\theta}_\infty) = \begin{pmatrix} -1 & 0 & 0 \\ 0 & -1 & 0 \\ \frac{\tau_1}{\tau_0^2} & -\frac{1}{\tau_0} & -\frac{1}{\tau_0} \end{pmatrix}. \quad (22)$$

Regarding $B(\boldsymbol{\theta}_\infty)$, elementary calculations show that

$$\psi(Z, \boldsymbol{\theta}_\infty) \psi(Z, \boldsymbol{\theta}_\infty)^T = \begin{pmatrix} (g_0(Z) - \tau_0)^2 & (g_0(Z) - \tau_0)(g_1(Z) - \tau_1) & 0 \\ (g_0(Z) - \tau_0)(g_1(Z) - \tau_1) & (g_1(Z) - \tau_1)^2 & 0 \\ 0 & 0 & 0 \end{pmatrix},$$

which leads to

$$B(\boldsymbol{\theta}_\infty) = \begin{pmatrix} \text{Var}[g_0(Z)] & \text{Cov}(g_0(Z), g_1(Z)) & 0 \\ \text{Cov}(g_0(Z), g_1(Z)) & \text{Var}[g_1(Z)] & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

Based on the previous calculations, we have

- $\psi(z, \boldsymbol{\theta})$ and its first two partial derivatives with respect to $\boldsymbol{\theta}$ exist for all z and for all $\boldsymbol{\theta}$ in the neighborhood of $\boldsymbol{\theta}_\infty$.
- For each $\boldsymbol{\theta}$ in the neighborhood of $\boldsymbol{\theta}_\infty$, we have for all $i, j, k \in \{0, 2\}$:

$$\left| \frac{\partial^2}{\partial \theta_i \partial \theta_j} \psi_k(z, \boldsymbol{\theta}) \right| \leq 1$$

and 1 is integrable.

- $A(\boldsymbol{\theta}_\infty)$ exists and is nonsingular.

- $B(\theta_\infty)$ exists and is finite.

Since we have:

$$\sum_{i=1}^n \psi(T_i, Y_i, \hat{\theta}_n) = 0 \quad \text{and} \quad \hat{\theta}_n \xrightarrow{P} \theta_\infty.$$

Then the conditions of Theorem 7.2 in Stefanski and Boos (2002) are satisfied, we have

$$\sqrt{n}(\hat{\theta}_n - \theta_\infty) \xrightarrow{d} \mathcal{N}(0, A(\theta_\infty)^{-1}B(\theta_\infty)(A(\theta_\infty)^{-1})^\top),$$

where

$$A(\theta_\infty)^{-1}B(\theta_\infty)(A(\theta_\infty)^{-1})^\top = \begin{bmatrix} \text{Var}[g_0(Z)] & \text{Cov}(g_0(Z), g_1(Z)) & \frac{\text{Cov}(g_0(Z), g_1(Z)) - \tau_1 \text{Var}[g_0(Z)]}{\tau_0} \\ \text{Cov}(g_0(Z), g_1(Z)) & \text{Var}[g_1(Z)] & -\frac{\text{Cov}(g_0(Z), g_1(Z))\tau_1}{\tau_0^2} + \frac{\text{Var}[g_1(Z)]}{\tau_0} \\ \frac{\text{Cov}(g_0(Z), g_1(Z)) - \tau_1 \text{Var}[g_0(Z)]}{\tau_0} & -\frac{\text{Cov}(g_0(Z), g_1(Z))\tau_1}{\tau_0^2} + \frac{\text{Var}[g_1(Z)]}{\tau_0} & V_{\text{RR}}^* \end{bmatrix}, \quad (23)$$

with

$$V_{\text{RR}}^* = \left(\frac{\tau_1}{\tau_0}\right)^2 \text{Var}\left(\frac{g_1(Z)}{\tau_1} - \frac{g_0(Z)}{\tau_0}\right). \quad (24)$$

In particular,

$$\sqrt{n}\left(\frac{\sum_{i=1}^n g_1(Z_i)}{\sum_{i=1}^n g_0(Z_i)} - \frac{\tau_1}{\tau_0}\right) \xrightarrow{d} \mathcal{N}(0, V_{\text{RR}}^*). \quad (25)$$

□

Theorem 2 (Finite sample bias and variance of the ratio of two estimators). *Let $T_1(\mathbf{Z})$ and $T_0(\mathbf{Z})$ be two unbiased estimators of τ_1 and $\tau_0 > 0$ where $\mathbf{Z} = (Z_1, \dots, Z_n)$ be n i.i.d. random variables. We assume that $M_0 \geq T_0(\mathbf{Z}) \geq m_0 > 0$, $|T_1(\mathbf{Z})| \leq M_1$. We also assume that $\text{Var}(T_1(\mathbf{Z})) = O_p(\frac{1}{n})$ and $\text{Var}(T_0(\mathbf{Z})) = O_p(\frac{1}{n})$. Then, we have that*

$$\text{Bias}\left(\frac{T_1(\mathbf{Z})}{T_0(\mathbf{Z})}, \frac{\tau_1}{\tau_0}\right) = \left| \mathbb{E}\left[\frac{T_1(\mathbf{Z})}{T_0(\mathbf{Z})}\right] - \frac{\tau_1}{\tau_0} \right| \leq \frac{M_1 M_0}{n m_0^2} \left(\frac{M_0}{m_0} + 1\right),$$

and

$$\left| \text{Var}\left(\frac{T_1(\mathbf{Z})}{T_0(\mathbf{Z})}\right) - \left(\frac{\tau_1}{\tau_0}\right)^2 \text{Var}\left(\frac{T_1(\mathbf{Z})}{\tau_1} - \frac{T_0(\mathbf{Z})}{\tau_0}\right) \right| \leq \frac{2M_0 M_1}{n m_0^4} \left(\frac{M_0 M_1}{m_0^2} + 1\right)$$

Proof. We rely on the multivariate version of Taylor's theorem to prove Theorem 2. We first introduce the multi-index notation:

$$|\alpha| = \alpha_1 + \dots + \alpha_n, \quad \alpha! = \alpha_1! \dots \alpha_n!, \quad \mathbf{x}^\alpha = x_1^{\alpha_1} \dots x_n^{\alpha_n}$$

and

$$D^\alpha f = \frac{\partial^{|\alpha|} f}{\partial x_1^{\alpha_1} \dots \partial x_n^{\alpha_n}}, \quad |\alpha| \leq k$$

Let f be the ratio function

$$f: \mathbb{R}_+^* \times \mathbb{R}_+^* \longrightarrow \mathbb{R} \\ (x_1, x_2) \longmapsto x_1/x_2.$$

Since f is two times continuously differentiable then one can derive an exact formula for the remainder in terms of 2nd order partial derivatives of f . Namely, if we define $\mathbf{x} = (x_1, x_2)$ and for $\mathbf{a} \in \mathbb{R}_+^* \times \mathbb{R}_+^*$

$$f(\mathbf{x}) = \sum_{|\alpha| \leq 1} \frac{D^\alpha f(\mathbf{a})}{\alpha!} (\mathbf{x} - \mathbf{a})^\alpha + R_{k+1}(\mathbf{x}), \quad (26)$$

with

$$R_{k+1}(\mathbf{x}) = \sum_{|\beta|=k+1} (\mathbf{x} - \mathbf{a})^\beta \frac{|\beta|}{\beta!} \int_0^1 (1-t)^{|\beta|-1} D^\beta f(\mathbf{a} + t(\mathbf{x} - \mathbf{a})) dt.$$

Bias:

Computing 26 for the ratio function with $\mathbf{x} = (T_1(\mathbf{Z}), T_0(\mathbf{Z}))$, $\mathbf{a} = (\tau_1, \tau_0)$ and taking the expectation gives us:

$$\begin{aligned} & \mathbb{E}[f(T_1(\mathbf{Z}), T_0(\mathbf{Z}))] \\ &= \mathbb{E} \left[f(\tau_1, \tau_0) + \frac{\partial f(\tau_1, \tau_0)}{\partial T_1(\mathbf{Z})} (T_1(\mathbf{Z}) - \tau_1) + \frac{\partial f(\tau_1, \tau_0)}{\partial T_0(\mathbf{Z})} (T_0(\mathbf{Z}) - \tau_0) + R_2(T_1(\mathbf{Z}), T_0(\mathbf{Z})) \right] \\ &= \mathbb{E}[f(\tau_1, \tau_0)] + \frac{\partial f(\tau_1, \tau_0)}{\partial T_1(\mathbf{Z})} \mathbb{E}[(T_1(\mathbf{Z}) - \tau_1)] \\ &+ \frac{\partial f(\tau_1, \tau_0)}{\partial T_0(\mathbf{Z})} \mathbb{E}[(T_0(\mathbf{Z}) - \tau_0)] + \mathbb{E}[R_2(T_1(\mathbf{Z}), T_0(\mathbf{Z}))] \\ &= \frac{\tau_1}{\tau_0} + \mathbb{E}[R_2(T_1(\mathbf{Z}), T_0(\mathbf{Z}))] \end{aligned}$$

In order to produce Theorem 2, we just need to show that $\mathbb{E}[R_2(T_1(\mathbf{Z}), T_0(\mathbf{Z}))] = O_p\left(\frac{1}{n}\right)$. To do so, we first compute $R_2(T_1(\mathbf{Z}), T_0(\mathbf{Z}))$

$$\begin{aligned} R_2(T_1(\mathbf{Z}), T_0(\mathbf{Z})) &= \underbrace{2(T_0(\mathbf{Z}) - \tau_0)^2 \int_0^1 \frac{(1-t)(\tau_1 + t(T_1(\mathbf{Z}) - \tau_1))}{(\tau_0 + t(T_0(\mathbf{Z}) - \tau_0))^3} dt}_{R_2^1(T_1(\mathbf{Z}), T_0(\mathbf{Z}))} \\ &- \underbrace{2(T_0(\mathbf{Z}) - \tau_0)(T_1(\mathbf{Z}) - \tau_1) \int_0^1 \frac{1-t}{(\tau_0 + t(T_0(\mathbf{Z}) - \tau_0))^2} dt}_{R_2^2(T_1(\mathbf{Z}), T_0(\mathbf{Z}))} \end{aligned}$$

Since we assume that $T_0(\mathbf{Z}) \geq m_0 > 0$ and that $|T_1(\mathbf{Z})| \leq M_1$ we have:

$$\begin{aligned} |R_2^1(T_1(\mathbf{Z}), T_0(\mathbf{Z}))| &= \left| 2(T_0(\mathbf{Z}) - \tau_0)^2 \int_0^1 \frac{(1-t)(\tau_1 + t(T_1(\mathbf{Z}) - \tau_1))}{(\tau_0 + t(T_0(\mathbf{Z}) - \tau_0))^3} dt \right| \\ &\leq 2(T_0(\mathbf{Z}) - \tau_0)^2 \int_0^1 \left| \frac{(1-t) \max(\tau_1, M_1)}{\min(m_0, \tau_0)^3} \right| dt \\ &= (T_0(\mathbf{Z}) - \tau_0)^2 \underbrace{\frac{M_1}{m_0^3}}_{C_1} \end{aligned}$$

Similarly, we have:

$$\begin{aligned} |R_2^2(T_1(\mathbf{Z}), T_0(\mathbf{Z}))| &= \left| 2(T_0(\mathbf{Z}) - \tau_0)(T_1(\mathbf{Z}) - \tau_1) \int_0^1 \frac{1-t}{(\tau_0 + t(T_0(\mathbf{Z}) - \tau_0))^2} dt \right| \\ &= 2|(T_0(\mathbf{Z}) - \tau_0)(T_1(\mathbf{Z}) - \tau_1)| \left| \int_0^1 \frac{1-t}{(\tau_0 + t(T_0(\mathbf{Z}) - \tau_0))^2} dt \right| \\ &\leq |(T_0(\mathbf{Z}) - \tau_0)(T_1(\mathbf{Z}) - \tau_1)| \underbrace{\frac{1}{m_0^2}}_{C_2} \end{aligned}$$

Finally we get that:

$$\begin{aligned} |\mathbb{E}[R_2(T_1(\mathbf{Z}), T_0(\mathbf{Z}))]| &\leq \mathbb{E}[|R_2^1(T_1(\mathbf{Z}), T_0(\mathbf{Z}))| + |R_2^2(T_1(\mathbf{Z}), T_0(\mathbf{Z}))|] \\ &\leq C_1 \text{Var}(T_0(\mathbf{Z})) + C_2 \mathbb{E}[|(T_0(\mathbf{Z}) - \tau_0)(T_1(\mathbf{Z}) - \tau_1)|] \\ &\leq C_1 \text{Var}(T_0(\mathbf{Z})) + C_2 \sqrt{\text{Var}(T_0(\mathbf{Z})) \text{Var}(T_1(\mathbf{Z}))} \\ &\leq C_1 M_0^2 + C_2 M_0 M_1 \end{aligned}$$

Since we have that $\text{Var}(T_1(\mathbf{Z})) = O_p\left(\frac{1}{n}\right)$ and $\text{Var}(T_0(\mathbf{Z})) = O_p\left(\frac{1}{n}\right)$, we can conclude by:

$$\text{Bias}\left(\frac{T_1(\mathbf{Z})}{T_0(\mathbf{Z})}, \frac{\tau_1}{\tau_0}\right) = \left| \mathbb{E}\left[\frac{T_1(\mathbf{Z})}{T_0(\mathbf{Z})}\right] - \frac{\tau_1}{\tau_0} \right| \lesssim \frac{M_1 M_0}{n m_0^2} \left(\frac{M_0}{m_0} + 1\right)$$

Variance:

Let us begin by expanding the variance of the function f :

$$\text{Var} f(T_1(\mathbf{Z}), T_0(\mathbf{Z})) = \mathbb{E} \left[(f(T_1(\mathbf{Z}), T_0(\mathbf{Z})) - \mathbb{E}[f(T_1(\mathbf{Z}), T_0(\mathbf{Z}))])^2 \right]$$

Next, apply Taylor's expansion around the means τ_1 and τ_0 :

$$\begin{aligned} &= \mathbb{E} \left[(f(\tau_1, \tau_0) + \frac{\partial f(\tau_1, \tau_0)}{\partial T_1(\mathbf{Z})} (T_1(\mathbf{Z}) - \tau_1) + \frac{\partial f(\tau_1, \tau_0)}{\partial T_0(\mathbf{Z})} (T_0(\mathbf{Z}) - \tau_0) \right. \\ &\quad \left. + R_2(T_1(\mathbf{Z}), T_0(\mathbf{Z})) - \mathbb{E}[f(T_1(\mathbf{Z}), T_0(\mathbf{Z}))])^2 \right] \end{aligned}$$

Simplify by focusing on the first-order derivatives and residual terms:

$$\begin{aligned} &= \mathbb{E} \left[\left(\frac{\partial f(\tau_1, \tau_0)}{\partial T_1(\mathbf{Z})} (T_1(\mathbf{Z}) - \tau_1) + \frac{\partial f(\tau_1, \tau_0)}{\partial T_0(\mathbf{Z})} (T_0(\mathbf{Z}) - \tau_0) \right. \right. \\ &\quad \left. \left. + R_2(T_1(\mathbf{Z}), T_0(\mathbf{Z})) - \mathbb{E}[R_2(T_1(\mathbf{Z}), T_0(\mathbf{Z}))]) \right)^2 \right] \end{aligned}$$

Decompose the variance into linear, cross-term, and residual contributions:

$$\begin{aligned} &= \mathbb{E} \left[\left(\frac{\partial f(\tau_1, \tau_0)}{\partial T_1(\mathbf{Z})} (T_1(\mathbf{Z}) - \tau_1) + \frac{\partial f(\tau_1, \tau_0)}{\partial T_0(\mathbf{Z})} (T_0(\mathbf{Z}) - \tau_0) \right)^2 \right] \\ &+ 2\mathbb{E} \left[\frac{\partial f(\tau_1, \tau_0)}{\partial T_1(\mathbf{Z})} (T_1(\mathbf{Z}) - \tau_1) \right. \\ &\quad \left. + \frac{\partial f(\tau_1, \tau_0)}{\partial T_0(\mathbf{Z})} (T_0(\mathbf{Z}) - \tau_0) \right] \mathbb{E} [R_2(T_1(\mathbf{Z}), T_0(\mathbf{Z})) - \mathbb{E}[R_2(T_1(\mathbf{Z}), T_0(\mathbf{Z}))]] \\ &+ \text{Var}(R_2(T_1(\mathbf{Z}), T_0(\mathbf{Z}))) \end{aligned}$$

Finally, re-express the result using a simplified ratio of variances:

$$= \left(\frac{\tau_1}{\tau_0}\right)^2 \text{Var}\left(\frac{T_1(\mathbf{Z})}{\tau_1} - \frac{T_0(\mathbf{Z})}{\tau_0}\right) + \text{Var}(R_2(T_1(\mathbf{Z}), T_0(\mathbf{Z})))$$

We now focus on $\text{Var}(R_2(T_1(\mathbf{Z}), T_0(\mathbf{Z})))$:

$$\begin{aligned} \text{Var}(R_2(T_1(\mathbf{Z}), T_0(\mathbf{Z}))) &\leq \mathbb{E} [(R_2(T_1(\mathbf{Z}), T_0(\mathbf{Z})))^2] \\ &\leq 2\mathbb{E} [|R_2^1(T_1(\mathbf{Z}), T_0(\mathbf{Z}))|^2] + 2\mathbb{E} [|R_2^2(T_1(\mathbf{Z}), T_0(\mathbf{Z}))|^2] \end{aligned}$$

We first focus on the first term:

$$\begin{aligned} \mathbb{E} [|R_2^1(T_1(\mathbf{Z}), T_0(\mathbf{Z}))|^2] &\leq \mathbb{E} \left[(C_{02}(T_0(\mathbf{Z}) - \tau_0)^2)^2 \right] \\ &\leq C_1^2 \mathbb{E} [(T_0(\mathbf{Z}) - \tau_0)^4] \\ &\leq C_1^2 M_0^2 \mathbb{E} [(T_0(\mathbf{Z}) - \tau_0)^2] \quad T_0(\mathbf{Z}) \leq M_0 \\ &\leq C_1^2 M_0^2 \text{Var}(T_0(\mathbf{Z})) \end{aligned}$$

For the second term we have:

$$\begin{aligned} \mathbb{E} [|R_2^2(T_1(\mathbf{Z}), T_0(\mathbf{Z}))|^2] &= C_2^2 \mathbb{E} [(T_0(\mathbf{Z}) - \tau_0)^2 (T_1(\mathbf{Z}) - \tau_1)^2] \\ &\leq C_2^2 \sqrt{\mathbb{E} [(T_0(\mathbf{Z}) - \tau_0)^4] \mathbb{E} [(T_1(\mathbf{Z}) - \tau_1)^4]} \\ T_1(\mathbf{Z}), T_0(\mathbf{Z}) \text{ bounded} &\leq C_2^2 M_0 M_1 \sqrt{\mathbb{E} [(T_0(\mathbf{Z}) - \tau_0)^2] \mathbb{E} [(T_1(\mathbf{Z}) - \tau_1)^2]} \\ &\leq C_2^2 M_0 M_1 \sqrt{\text{Var}(T_0(\mathbf{Z})) \text{Var}(T_1(\mathbf{Z}))} \end{aligned}$$

Hence we get that:

$$\text{Var}(R_2(T_1(\mathbf{Z}), T_0(\mathbf{Z}))) \lesssim \frac{2M_0 M_1}{n m_0^4} \left(\frac{M_0 M_1}{m_0^2} + 1\right)$$

□

This theorem is an adaptation of Theorem 4.1 of Derumigny et al. (2019).

A.1 Proofs of within-trial estimators of the risk ratio

A.1.1 Ratio Horvitz-Thomson estimator

Proof of Proposition 1.

Asymptotic Bias and Variance: Let $Z_i := (T_i, Y_i)$ and define $g_0(Z_i) = \frac{(1-T_i)Y_i}{1-e}$ and $g_1(Z_i) = \frac{T_i Y_i}{e}$. First, we evaluate the expectation of $g_1(Z_i)$:

$$\begin{aligned}
\mathbb{E}[g_1(Z_i)] &= \mathbb{E}\left[\frac{T_i Y_i}{e}\right] && \text{(by i.i.d)} \\
&= \mathbb{E}\left[\frac{T_i Y_i^{(1)}}{e}\right] && \text{(by SUTVA)} \\
&= \mathbb{E}\left[\frac{T_i}{e}\right] \mathbb{E}[Y_i^{(1)}] && \text{(by ignorability)} \\
&= \mathbb{E}[Y_i^{(1)}] && \text{(by Trial positivity)}
\end{aligned}$$

Similarly, we can find the expectation of $g_0(Z_i)$:

$$\mathbb{E}[g_0(Z_i)] = \mathbb{E}[Y^{(0)}] > 0.$$

Thus, according to Theorem 1, we have

$$\sqrt{n}(\hat{\tau}_{\text{RR-HT}, n} - \tau_{\text{RR}}) \xrightarrow{d} \mathcal{N}(0, V_{\text{RR-HT}}),$$

with

$$\begin{aligned}
V_{\text{RR-HT}} &= \left(\frac{\tau_1}{\tau_0}\right)^2 \text{Var}\left(\frac{g_1(Z)}{\tau_1} - \frac{g_0(Z)}{\tau_0}\right) \\
&= \tau_{\text{RR}}^2 \text{Var}\left(\frac{TY}{e\mathbb{E}[Y^{(1)}]} - \frac{(1-T)Y}{(1-e)\mathbb{E}[Y^{(0)}]}\right).
\end{aligned}$$

Next, we evaluate the variance terms separately:

$$\begin{aligned}
\text{Var}\left(\frac{TY}{e\mathbb{E}[Y^{(1)}]}\right) &= \frac{1}{\mathbb{E}[Y^{(1)}]^2 e^2} \text{Var}(TY) \\
&= \frac{1}{\mathbb{E}[Y^{(1)}]^2 e^2} \left(\mathbb{E}[(TY)^2] - \mathbb{E}[TY]^2\right) \\
&= \frac{1}{\mathbb{E}[Y^{(1)}]^2 e^2} \left(\mathbb{E}[T(Y)^2] - \mathbb{E}[TY]^2\right) && (T \text{ is binary}) \\
&= \frac{1}{\mathbb{E}[Y^{(1)}]^2 e^2} \left(\mathbb{E}[T(Y^{(1)})^2] - \mathbb{E}[TY^{(1)}]^2\right) && \text{(by SUTVA)} \\
&= \frac{1}{\mathbb{E}[Y^{(1)}]^2 e^2} \left(e\mathbb{E}[(Y^{(1)})^2] - e^2\mathbb{E}[Y^{(1)}]^2\right) && \text{(by ignorability)} \\
&= \frac{\mathbb{E}[(Y^{(1)})^2]}{e\mathbb{E}[Y^{(1)}]^2} - 1.
\end{aligned}$$

Similarly, we find the variance of the second term:

$$\text{Var} \left(\frac{(1-T)Y}{(1-e)\mathbb{E}[Y^{(0)}]} \right) = \frac{\mathbb{E}[(Y^{(0)})^2]}{(1-e)\mathbb{E}[Y^{(0)}]^2} - 1.$$

Finally, we compute the covariance between the two terms:

$$\begin{aligned} \text{Cov} \left(\frac{TY}{e\mathbb{E}[Y^{(1)}]}, \frac{(1-T)Y}{(1-e)\mathbb{E}[Y^{(0)}]} \right) &= \frac{\text{Cov}(TY, (1-T)Y)}{e\mathbb{E}[Y^{(1)}] (1-e)\mathbb{E}[Y^{(0)}]} \\ &= \frac{(\mathbb{E}[T(1-T)Y^2] - \mathbb{E}[TY]\mathbb{E}[(1-T)Y])}{e\mathbb{E}[Y^{(1)}] (1-e)\mathbb{E}[Y^{(0)}]} \\ &= \frac{-\mathbb{E}[TY]\mathbb{E}[(1-T)Y]}{e\mathbb{E}[Y^{(1)}] (1-e)\mathbb{E}[Y^{(0)}]} \\ &= -1. \end{aligned}$$

Using Bienaymé's identity, we finally obtain:

$$V_{\text{RR-HT}} = \tau_{\text{RR}}^2 \left(\frac{\mathbb{E}[(Y^{(1)})^2]}{e\mathbb{E}[Y^{(1)}]^2} + \frac{\mathbb{E}[(Y^{(0)})^2]}{(1-e)\mathbb{E}[Y^{(0)}]^2} \right).$$

Finite sample Bias and Variance: Let $T_1(\mathbf{Z}) = \frac{1}{n} \sum_{i=1}^n \frac{T_i Y_i}{e}$ and $T_0(\mathbf{Z}) = \frac{1}{n} \sum_{i=1}^n \frac{(1-T_i)Y_i}{1-e}$ where $\mathbf{Z} = (Z_1, \dots, Z_n)$. Note that we have:

First, consider the variance of $T_1(\mathbf{Z})$:

$$\begin{aligned} \text{Var}(T_1(\mathbf{Z})) &= \frac{1}{ne^2} \text{Var}(T_i Y_i) && \text{(by i.i.d)} \\ &= \frac{1}{ne^2} \left(\mathbb{E}[(T_i Y_i)^2] - \mathbb{E}[T_i Y_i]^2 \right) \\ &= \frac{\mathbb{E}[(Y^{(1)})^2] - e\mathbb{E}[Y^{(1)}]^2}{ne} = O_p \left(\frac{1}{n} \right). \end{aligned}$$

Similarly, we also have that:

$$\text{Var}(T_0(\mathbf{Z})) = O_p \left(\frac{1}{n} \right).$$

Next, we show that $T_0(\mathbf{Z})$ is bounded:

$$\begin{aligned}
T_0(\mathbf{Z}) &= \frac{1}{n} \sum_{i=1}^n \frac{(1-T_i)Y_i}{1-e} \\
&= \frac{1}{n(1-e)} \sum_{i=1}^n (1-T_i)Y_i \\
&\geq \frac{m}{(1-e)} \sum_{i=1}^n (1-T_i) && \text{(since } Y_i \geq m > 0\text{)} \\
&\geq \frac{m}{(1-e)} && \text{(as } \sum_{i=1}^n T_i < n\text{)}.
\end{aligned}$$

Similarly, we also have the upper bound:

$$\begin{aligned}
T_0(\mathbf{Z}) &= \frac{1}{n} \sum_{i=1}^n \frac{(1-T_i)Y_i}{1-e} \\
&= \frac{1}{ne} \sum_{i=1}^n (1-T_i)Y_i \\
&\leq \frac{1}{ne} \sum_{i=1}^n Y_i && \text{(since } T \text{ is binary)} \\
&\leq \frac{M}{e} && \text{(since } Y_i \leq M\text{)}.
\end{aligned}$$

Similarly, for $T_1(\mathbf{Z})$, we have:

$$T_1(\mathbf{Z}) \leq \frac{M}{e}$$

Therefore, we have shown that $T_1(\mathbf{Z})$ and $T_0(\mathbf{Z})$ are unbiased estimators of $\mathbb{E}[Y^{(1)}]$ and $\mathbb{E}[Y^{(0)}] > 0$, respectively. We also established that $M/e \geq T_0(\mathbf{Z}) \geq m/(1-e) > 0$ and $|T_1(\mathbf{Z})| \leq M/e$.

Furthermore, we pointed out that $\text{Var}(T_1(\mathbf{Z})) = O_p(\frac{1}{n})$ and $\text{Var}(T_0(\mathbf{Z})) = O_p(\frac{1}{n})$. Applying Theorem 2, we obtain:

$$|\mathbb{E}[\hat{\tau}_{\text{RR, HT, n}}] - \tau_{\text{RR}}| \leq \frac{M^2(1-e)^2}{ne^2m^2} \left(\frac{M(1-e)}{me} + 1 \right) \leq \frac{2M^3(1-e)^3}{nm^3e^3}$$

and

$$|\text{Var}(\hat{\tau}_{\text{RR, HT, n}}) - V_{\text{RR, HT}}| \leq \frac{2M^2(1-e)^4}{nm^4e^2} \left(\frac{M^2(1-e)^2}{m^2e^2} + 1 \right) \leq \frac{4M^4(1-e)^6}{nm^6e^4}$$

Optimal choice of e : the optimal value of e_{opt} is the one that minimizes the variance of the Ratio Horvitz-Thomson estimator. Therefore, we need to solve:

$$\inf_{e \in (0,1)} \tau_{\text{RR}}^2 \left(\frac{\mathbb{E}[(Y^{(1)})^2]}{e\mathbb{E}[Y^{(1)}]^2} + \frac{\mathbb{E}[(Y^{(0)})^2]}{(1-e)\mathbb{E}[Y^{(0)}]^2} \right)$$

Noting that the variance we want to minimize is convex in e , we can derive the variance and set it to 0 to find e_{opt} . We have:

$$\frac{C_1}{e_{\text{opt}}^2} = \frac{C_0}{(1-e_{\text{opt}})^2}$$

where $C_1 = \frac{\mathbb{E}[(Y^{(1)})^2]}{\mathbb{E}[Y^{(1)}]^2}$ and $C_0 = \frac{\mathbb{E}[(Y^{(0)})^2]}{\mathbb{E}[Y^{(0)}]^2}$.

- If $\frac{\text{Var}(Y^{(1)})}{\mathbb{E}[Y^{(1)}]^2} = \frac{\text{Var}(Y^{(0)})}{\mathbb{E}[Y^{(0)}]^2}$:

$$e_{opt} = 0.5$$

- otherwise:

$$e_{opt} = \frac{\mathbb{E}[(Y^{(1)})^2] \mathbb{E}[Y^{(0)}]^2 - \sqrt{\mathbb{E}[(Y^{(1)})^2] \mathbb{E}[(Y^{(0)})^2]} \mathbb{E}[Y^{(1)}] \mathbb{E}[Y^{(0)}]}{\mathbb{E}[(Y^{(1)})^2] \mathbb{E}[Y^{(0)}]^2 - \mathbb{E}[(Y^{(0)})^2] \mathbb{E}[Y^{(1)}]^2} \in (0, 1)$$

□

A.1.2 Risk Ratio Neyman estimator

Proof of Proposition 2.

Asymptotic Bias and Variance: we proceed with M-estimations to prove asymptotic bias and variance of the Ratio Neyman estimator, we first define the following:

$$\hat{\theta}_n = \begin{pmatrix} \frac{1}{n_0} \sum_{T_i=0} Y_i \\ \frac{1}{n_1} \sum_{T_i=1} Y_i \\ \hat{\tau}_{R-N,n} \end{pmatrix} \quad \text{and} \quad \psi(T, Y, \theta) = \begin{pmatrix} \psi_0(\theta) \\ \psi_1(\theta) \\ \psi_2(\theta) \end{pmatrix} =: \begin{pmatrix} (1-T)(Y - \theta_0) \\ T(Y - \theta_1) \\ \theta_1 - \theta_2 \theta_0 \end{pmatrix}, \quad (27)$$

where $\theta = (\theta_0, \theta_1, \theta_2)$.

Next, we verify that for $\hat{\theta}_n = (\frac{1}{n_0} \sum_{T_i=0} Y_i, \frac{1}{n_1} \sum_{T_i=1} Y_i, \hat{\tau}_{R-N,n})$, we have:

$$\sum_{i=1}^n \psi(T_i, Y_i, \hat{\theta}_n) = 0.$$

We begin by demonstrating this for ψ_1 :

$$\begin{aligned} \sum_{i=1}^n \psi_1(T_i, Y_i, \hat{\theta}_n) &= \sum_{i=1}^n T_i \left(Y_i - \frac{1}{n_1} \sum_{T_j=1} Y_j \right) \\ &= \sum_{i=1}^n T_i \left(Y_i - \frac{1}{n_1} \sum_{j=1}^n T_j Y_j \right) \\ &= \sum_{i=1}^n T_i Y_i - \frac{1}{n_1} \sum_{i=1}^n T_i \sum_{j=1}^n T_j Y_j \\ &= \sum_{i=1}^n T_i Y_i - \sum_{j=1}^n T_j Y_j \\ &= 0. \end{aligned}$$

Similarly, we can show:

$$\sum_{i=1}^n \psi_0(T_i, Y_i, \hat{\theta}_n) = 0.$$

Moreover, by construction:

$$\sum_{i=1}^n \psi_2(T_i, Y_i, \hat{\theta}_n) = 0.$$

Thus, we have established that $\hat{\theta}_n$ is an M-estimator of type ψ (see Stefanski and Boos, 2002). Given that we are in a Bernoulli Trial, we now demonstrate that $\mathbb{E}[\psi(T, Y, \theta_\infty)] = 0$ where $\theta_\infty =$

$(\mathbb{E}[Y^{(0)}], \mathbb{E}[Y^{(1)}], \tau_{RR})$. Therefore, we have:

$$\begin{aligned}\mathbb{E}[\psi_1(\theta_\infty)] &= \mathbb{E}\left[T\left(Y - \mathbb{E}[Y^{(1)}]\right)\right] \\ &= \mathbb{E}\left[T\left(Y^{(1)} - \mathbb{E}[Y^{(1)}]\right)\right] && \text{(by SUTVA)} \\ &= \mathbb{E}[T]\mathbb{E}\left[Y^{(1)} - \mathbb{E}[Y^{(1)}]\right] && \text{(by ignorability)} \\ &= 0.\end{aligned}$$

Similarly, we can show:

$$\mathbb{E}[\psi_0(\theta_\infty)] = 0.$$

Furthermore, we have:

$$\mathbb{E}[\psi_2(\theta_\infty)] = \mathbb{E}[Y^{(1)}] - \tau_{RR}\mathbb{E}[Y^{(0)}] = 0.$$

At this point, we note that θ_∞ is the only value of θ such that $\mathbb{E}[\psi(T, Y, \theta)] = 0$. We proceed by defining:

$$A(\theta_\infty) = \mathbb{E}\left[\frac{\partial\psi}{\partial\theta}\Big|_{\theta=\theta_\infty}\right] \quad \text{and} \quad B(\theta_\infty) = \mathbb{E}[\psi(Z, \theta_\infty)\psi(Z, \theta_\infty)^T].$$

Next, we check the conditions of Theorem 7.2 in Stefanski and Boos (2002). First, we compute $A(\theta_\infty)$ and $B(\theta_\infty)$. Since:

$$\frac{\partial\psi}{\partial\theta}(Z, \theta) = \begin{pmatrix} -(1-T) & 0 & 0 \\ 0 & -T & 0 \\ -\theta_2 & 1 & -\theta_0 \end{pmatrix},$$

we obtain:

$$A(\theta_\infty) = \begin{pmatrix} -(1-e) & 0 & 0 \\ 0 & -e & 0 \\ -\tau_{RR} & 1 & -\mathbb{E}[Y^{(0)}] \end{pmatrix},$$

which leads to:

$$A^{-1}(\theta_\infty) = \begin{pmatrix} \frac{1}{e-1} & 0 & 0 \\ 0 & -\frac{1}{e} & 0 \\ \tau_{RR}\frac{1}{\mathbb{E}[Y^{(0)}](1-e)} & -\frac{1}{e\mathbb{E}[Y^{(0)}]} & -\frac{1}{\mathbb{E}[Y^{(0)}]} \end{pmatrix}.$$

Regarding $B(\theta_\infty)$, elementary calculations show that:

$$\begin{aligned}&\psi(Z, \theta_\infty)\psi(Z, \theta_\infty)^T \\ &= \begin{pmatrix} ((1-T)(Y - \mathbb{E}[Y^{(0)}]))^2 & (1-T)(Y - \mathbb{E}[Y^{(0)}])T(Y - \mathbb{E}[Y^{(1)}]) & 0 \\ (1-T)(Y - \mathbb{E}[Y^{(0)}])T(Y - \mathbb{E}[Y^{(1)}]) & (T(Y - \mathbb{E}[Y^{(1)}]))^2 & 0 \\ 0 & 0 & 0 \end{pmatrix},\end{aligned}$$

which leads to:

$$B(\theta_\infty) = \begin{pmatrix} (1-e)\text{Var}[Y^{(0)}] & 0 & 0 \\ 0 & e\text{Var}[Y^{(1)}] & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

Based on the previous calculations, we have:

- $\psi(z, \theta)$ and its first two partial derivatives with respect to θ exist for all z and for all θ in the neighborhood of θ_∞ .
- For each θ in the neighborhood of θ_∞ , we have for all $i, j, k \in \{0, 2\}$:

$$\left|\frac{\partial^2}{\partial\theta_i\partial\theta_j}\psi_k(z, \theta)\right| \leq 1$$

and 1 is integrable.

- $A(\theta_\infty)$ exists and is nonsingular.
- $B(\theta_\infty)$ exists and is finite.

Since we have:

$$\sum_{i=1}^n \psi(T_i, Y_i, \hat{\theta}_n) = 0 \quad \text{and} \quad \hat{\theta}_n \xrightarrow{p} \theta_\infty.$$

Then, the conditions of Theorem 7.2 in Stefanski and Boos (2002) are satisfied, we have:

$$\sqrt{n} (\hat{\theta}_n - \theta_\infty) \xrightarrow{d} \mathcal{N} (0, A(\theta_\infty)^{-1} B(\theta_\infty) (A(\theta_\infty)^{-1})^\top),$$

where:

$$A(\theta_\infty)^{-1} B(\theta_\infty) (A(\theta_\infty)^{-1})^\top = \begin{bmatrix} \frac{\text{Var}[Y^{(0)}]}{(1-e)} & 0 & -\frac{\tau \text{Var}[Y^{(0)}]}{\tau_0(1-e)} \\ 0 & \frac{\text{Var}[Y^{(1)}]}{e} & \frac{\text{Var}[Y^{(1)}]}{e\tau_0} \\ -\frac{\tau \text{Var}[Y^{(0)}]}{\tau_0(1-e)} & \frac{\text{Var}[Y^{(1)}]}{e\tau_0} & V_{R-N} \end{bmatrix},$$

with:

$$V_{R-N} = \tau_{RR}^2 \left(\frac{\text{Var}(Y^{(1)})}{e\mathbb{E}[Y^{(1)}]^2} + \frac{\text{Var}(Y^{(0)})}{(1-e)\mathbb{E}[Y^{(0)}]^2} \right).$$

In particular, we obtain:

$$\sqrt{n} (\hat{\tau}_{RR,n} - \tau_{RR}) \xrightarrow{d} \mathcal{N} (0, V_{RR,N}).$$

Finally, note that:

$$\begin{aligned} V_{R-N} &= \tau_{RR}^2 \left(\frac{\text{Var}(Y^{(1)})}{e\mathbb{E}[Y^{(1)}]^2} + \frac{\text{Var}(Y^{(0)})}{(1-e)\mathbb{E}[Y^{(0)}]^2} \right) \\ &= \tau_{RR}^2 \left(\frac{\mathbb{E}[(Y^{(1)})^2] - \mathbb{E}[Y^{(1)}]^2}{e\mathbb{E}[Y^{(1)}]^2} + \frac{\mathbb{E}[(Y^{(0)})^2] - \mathbb{E}[Y^{(0)}]^2}{(1-e)\mathbb{E}[Y^{(0)}]^2} \right) \\ &= V_{R-HT} - \frac{\tau_{RR}^2}{e(1-e)}. \end{aligned}$$

Finite sample bias and variance: let $T_1(\mathbf{Z}) = \frac{1}{n} \sum_{i=1}^n \frac{T_i Y_i}{\hat{e}}$ and $T_0(\mathbf{Z}) = \frac{1}{n} \sum_{i=1}^n \frac{(1-T_i) Y_i}{1-\hat{e}}$ where $\mathbf{Z} = (Z_1, \dots, Z_n)$ and $\hat{e} = n_1/n$. We first show that $\text{Var}(T_1(\mathbf{Z})) = O_p(\frac{1}{n})$ and $\text{Var}(T_0(\mathbf{Z})) = O_p(\frac{1}{n})$. This part is inspired by Colnet et al. (2024a):

Using the Eve's law, and conditioning on the treatment assignment vector $\mathbf{T} = (T_1, \dots, T_n)$, one has

$$\text{Var}(T_1(\mathbf{Z})) = \text{Var}(\mathbb{E}[T_1(\mathbf{Z}) | \mathbf{T}]) + \mathbb{E}[\text{Var}(T_1(\mathbf{Z}) | \mathbf{T})].$$

Since we also have that:

$$\begin{aligned} \mathbb{E}[T_1(\mathbf{Z}) | \mathbf{T}] &= \frac{\frac{1}{n} \sum_{i=1}^n T_i}{\frac{1}{n} \sum_{i=1}^n T_i} \mathbb{E}[Y_i^{(1)} | \mathbf{T}] \\ &= \mathbf{1}_{\sum_{i=1}^n T_i > 0} \mathbb{E}[Y_i^{(1)} | \mathbf{T}] \\ &= \mathbf{1}_{\sum_{i=1}^n T_i > 0} \mathbb{E}[Y_i^{(1)}] \quad \text{(by ignorability)} \end{aligned}$$

we can now compute the variance of this quantity:

$$\begin{aligned}\text{Var} [\mathbb{E} [T_1(\mathbf{Z}) \mid \mathbf{T}]] &= \text{Var} \left(\mathbf{1}_{\sum_{i=1}^n A_i > 0} \mathbb{E} [Y_i^{(1)}] \right) \\ &= \mathbb{E} [Y_i^{(1)}]^2 \text{Var} (\mathbf{1}_{\sum_{i=1}^n T_i > 0})\end{aligned}$$

Besides,

$$\begin{aligned}\text{Var} (\mathbf{1}_{\sum_{i=1}^n T_i > 0}) &= \mathbb{E} [\mathbf{1}_{\sum_{i=1}^n T_i > 0}^2] - \mathbb{E} [\mathbf{1}_{\sum_{i=1}^n T_i > 0}]^2 \\ &= 1 - (1 - e)^n - (1 - (1 - e)^n)^2 \\ &= (1 - e)^n (1 - (1 - e)^n),\end{aligned}$$

Therefore, we get:

$$\begin{aligned}\text{Var} [\mathbb{E} [T_1(\mathbf{Z}) \mid \mathbf{T}]] &= \mathbb{E} [Y_i^{(1)}]^2 (1 - e)^n (1 - (1 - e)^n) \\ &\leq (1 - e)^n \mathbb{E} [Y_i^{(1)}]^2 = o_p \left(\frac{1}{n} \right)\end{aligned}$$

Now,

$$\begin{aligned}\text{Var} (T_1(\mathbf{Z}) \mid \mathbf{T}) &= \text{Var} \left(\frac{1}{n} \sum_{i=1}^n \frac{T_i Y_i^{(1)}}{\hat{e}} \mid \mathbf{T} \right) \\ &= \frac{1}{n} \text{Var} \left[\frac{T_i Y_i^{(1)}}{\hat{e}} \mid \mathbf{T} \right] && \text{(by i.i.d.)} \\ &= \frac{1}{n} \left(\frac{T_i}{\hat{e}} \right)^2 \text{Var} [Y_i^{(1)} \mid \mathbf{T}] \\ &= \frac{1}{n} \left(\frac{T_i}{\hat{e}} \right)^2 \text{Var} [Y_i^{(1)}] && \text{(by ignorability)}\end{aligned}$$

Taking the expectation of the previous term leads to,

$$\begin{aligned}\mathbb{E} [\text{Var} (T_1(\mathbf{Z}) \mid \mathbf{T})] &= \mathbb{E} \left[\frac{1}{n} \left(\frac{A_i}{\hat{e}} \right)^2 \text{Var} [Y_i^{(1)}] \right] \\ &= \frac{1}{n} \mathbb{E} \left[\left(\frac{A_i}{\hat{e}} \right)^2 \right] \text{Var} [Y_i^{(1)}] && \text{by linearity.}\end{aligned}$$

Note that,

$$\begin{aligned}\mathbb{E} \left[\left(\frac{T_i}{\hat{e}} \right)^2 \right] &= \mathbb{E} \left[\frac{T_i}{(\hat{e})^2} \right] \\ &= \frac{1}{n} \left(\mathbb{E} \left[\frac{T_1}{\hat{e}^2} \right] + \mathbb{E} \left[\frac{T_2}{\hat{e}^2} \right] + \dots + \mathbb{E} \left[\frac{T_n}{\hat{e}^2} \right] \right) \\ &= \mathbb{E} \left[\frac{\hat{e}}{\hat{e}^2} \right] \\ &= \mathbb{E} \left[\frac{\mathbf{1}_{\hat{e} > 0}}{\hat{e}} \right],\end{aligned}$$

so that

$$\mathbb{E}[\text{Var}(T_1(\mathbf{Z}) \mid \mathbf{T})] = \frac{1}{n} \mathbb{E} \left[\frac{\mathbf{1}_{\hat{e} > 0}}{\hat{e}} \right] \text{Var} \left[Y_i^{(1)} \right] = O_p \left(\frac{1}{n} \right).$$

Therefore we have that

$$\text{Var}(T_1(\mathbf{Z})) = O_p \left(\frac{1}{n} \right)$$

. Similarly we can show that

$$\text{Var}(T_0(\mathbf{Z})) = O_p \left(\frac{1}{n} \right)$$

Next, we show that $T_0(\mathbf{Z})$ is bounded

$$\begin{aligned} T_0(\mathbf{Z}) &= \frac{1}{n} \sum_{i=1}^n \frac{(1 - T_i) Y_i}{1 - \hat{e}} \\ &\geq \frac{m}{n} \sum_{i=1}^n \frac{1 - T_i}{1 - \hat{e}} && \text{(since } Y_i \geq m > 0) \\ &\geq m. && \text{(by definition of } \hat{e}) \end{aligned}$$

Similarly, we also have the upper bound:

$$\begin{aligned} T_0(\mathbf{Z}) &= \frac{1}{n} \sum_{i=1}^n \frac{(1 - T_i) Y_i}{1 - \hat{e}} \\ &\leq \frac{M}{n} \sum_{i=1}^n \frac{1 - T_i}{1 - \hat{e}} && \text{(since } Y_i \leq M) \\ &\leq M && \text{(by definition of } \hat{e}) \end{aligned}$$

Similarly, for $T_1(\mathbf{Z})$, we have:

$$\begin{aligned} T_1(\mathbf{Z}) &= \frac{1}{n} \sum_{i=1}^n \frac{T_i Y_i}{\hat{e}} \\ &\leq \frac{M}{n} \sum_{i=1}^n \frac{T_i}{\hat{e}} && \text{(since } Y_i \leq M) \\ &\leq M && \text{(by definition of } \hat{e}) \end{aligned}$$

Therefore, we have shown that $T_1(\mathbf{Z})$ and $T_0(\mathbf{Z})$ are unbiased estimators of $\mathbb{E}[Y^{(1)}]$ and $\mathbb{E}[Y^{(0)}] > 0$, respectively. We also established that $M \geq T_0(\mathbf{Z}) \geq m > 0$ and $|T_1(\mathbf{Z})| \leq M$.

Furthermore, we pointed out that $\text{Var}(T_1(\mathbf{Z})) = O_p \left(\frac{1}{n} \right)$ and $\text{Var}(T_0(\mathbf{Z})) = O_p \left(\frac{1}{n} \right)$. Applying Theorem 2, we obtain:

$$|\mathbb{E}[\hat{\tau}_{\text{RR, HT, n}}] - \tau_{\text{RR}}| \leq \frac{2M^3}{nm^3} \quad \text{and} \quad |\text{Var}(\hat{\tau}_{\text{RR, HT, n}}) - V_{\text{RR, HT}}| \leq \frac{4M^4}{nm^6}$$

Optimal choice of e : the optimal value of e_{opt} is the one that minimizes the variance of the Ratio Neyman estimator. Therefore, we need to solve:

$$\inf_{e \in (0,1)} \tau_{RR}^2 \left(\frac{\text{Var}(Y^{(1)})}{e \mathbb{E}[Y^{(1)}]^2} + \frac{\text{Var}(Y^{(0)})}{(1-e) \mathbb{E}[Y^{(0)}]^2} \right)$$

Noting that the variance we want to minimize is convex in e , we can derive the variance and set it to 0 to find e_{opt} . We have:

$$\frac{C_1}{e_{opt}^2} = \frac{C_0}{(1-e_{opt})^2}$$

where $C_1 = \frac{\text{Var}(Y^{(1)})}{\mathbb{E}[Y^{(1)}]^2}$ and $C_0 = \frac{\text{Var}(Y^{(0)})}{\mathbb{E}[Y^{(0)}]^2}$.

- If $\frac{\text{Var}(Y^{(1)})}{\mathbb{E}[Y^{(1)}]^2} = \frac{\text{Var}(Y^{(0)})}{\mathbb{E}[Y^{(0)}]^2}$:

$$e_{opt} = 0.5$$

- otherwise:

$$e_{opt} = \frac{C_1 - \sqrt{C_1 C_0}}{C_1 - C_0} \in (0, 1)$$

□

A.1.3 Ratio G formula estimator

Proof of Proposition 3.

Asymptotic bias and variance of the oracle ratio G formula estimator: we remind that the oracle ratio G formula is defined as:

$$\tau_{RR,G,n}^* = \frac{\sum_{i=1}^n \mu_1^*(X_i)}{\sum_{i=1}^n \mu_0^*(X_i)},$$

where the response surfaces μ_0^* and μ_1^* are assumed to be known. Let us define $g_1(Z) = \mu_1^*(X_i)$ and $g_0(Z) = \mu_0^*(X_i)$ with $Z = X$. Since both of the $g_1(Z)$ and $g_0(Z)$ are bounded, they are square integrable. We also have that $\mathbb{E}[g_0(Z_i)] = \mathbb{E}[Y^{(0)}]$ and $\mathbb{E}[g_1(Z_i)] = \mathbb{E}[Y^{(1)}]$. We can therefore apply Theorem 2 and conclude that:

$$\sqrt{n}(\tau_{RR,G,n}^* - \tau_{RR}) \rightarrow \mathcal{N}(0, V_{RR,G})$$

where $V_{RR,G} = \tau_{RR}^2 \text{Var} \left(\frac{\mu_1^*(X)}{\mathbb{E}[Y^{(1)}]} - \frac{\mu_0^*(X)}{\mathbb{E}[Y^{(0)}]} \right)$.

Finite sample bias and variance of the oracle ratio G formula estimator: let $T_1(\mathbf{Z}) = \frac{1}{n} \sum_{i=1}^n \mu_1^*(X_i)$ and $T_0(\mathbf{Z}) = \frac{1}{n} \sum_{i=1}^n \mu_0^*(X_i)$ where $\mathbf{Z} = (X_1, \dots, X_n)$. We first show that $\text{Var}(T_1(\mathbf{Z})) = O_p\left(\frac{1}{n}\right)$ and $\text{Var}(T_0(\mathbf{Z})) = O_p\left(\frac{1}{n}\right)$:

$$\begin{aligned} \text{Var}(T_1(\mathbf{Z})) &= \frac{1}{n^2} \text{Var} \left(\sum_{i=1}^n \mu_1^*(X_i) \right) \\ &= \frac{1}{n^2} \sum_{i=1}^n \text{Var}(\mu_1^*(X_i)) && \text{(by i.i.d.)} \\ &= \frac{1}{n} \left(\mathbb{E}[(\mu_1^*(X_i))^2] - \mathbb{E}[Y^{(1)}]^2 \right) && \text{(by law of total expectation)} \\ &\leq \frac{M^2 - \mathbb{E}[Y^{(1)}]^2}{n} && (\mu_1^*(X_i) \leq M) \\ &= O_p\left(\frac{1}{n}\right) \end{aligned}$$

Similarly,

$$\text{Var}(T_0(\mathbf{Z})) = O_p\left(\frac{1}{n}\right)$$

Therefore, we have shown that $T_1(\mathbf{Z})$ and $T_0(\mathbf{Z})$ are unbiased estimators of $\mathbb{E}[Y^{(1)}]$ and $\mathbb{E}[Y^{(0)}] > 0$, respectively. We also have that $T_0(\mathbf{Z})$ and $T_1(\mathbf{Z})$ are bounded:

$$m_0 \leq T_0(\mathbf{Z}) \leq M_0 \quad \text{and} \quad T_1(\mathbf{Z}) \leq M_1$$

Applying Theorem 2, we obtain:

$$|\mathbb{E}[\hat{\tau}_{\text{RR, HT, n}}] - \tau_{\text{RR}}| \leq \frac{2M_1M_0^2}{nm_0^3} \quad \text{and} \quad |\text{Var}(\hat{\tau}_{\text{RR, HT, n}}) - V_{\text{RR, HT}}| \leq \frac{2M_0^2M_1(M_1 + M_0)}{m_0^6}$$

□

A.1.4 RCT linear model estimation - Proof of Lemma 1

We use the linear model setting, introduced by Wager (2020). More precisely, we assume that, for all $t \in \{0, 1\}$,

$$Y_i^{(t)} = c^{(t)} + X_i^\top \beta^{(t)} + \varepsilon_i^{(t)}, \quad \mathbb{E}[\varepsilon_i^{(t)} | X_i] = 0, \quad \text{Var}[\varepsilon_i^{(t)} | X_i] = \sigma^2, \quad (28)$$

with $\mathbb{E}[X] = 0$. We define the covariance matrix of X as $\Sigma = \text{Var}[X]$, the corresponding dot product as $\langle X, Y \rangle_\Sigma = X^\top \Sigma Y$, and its corresponding norm is given by $\|X\|_\Sigma = \sqrt{\langle X, X \rangle_\Sigma}$. We also assume that $\|X\|_\Sigma \leq M$.

Lemma 1 (Asymptotic variance comparison of $\hat{\tau}_{\text{RR, N}}$ and $\hat{\tau}_{\text{RR, G}}$ under a linear model). *Grant Assumption 1 and Assumption 2. Within the linear model framework defined above and using linear regression models for both of the surface responses, we have $V_{\text{RR, N}} > V_{\text{RR, G}}$, provided that either $\beta^{(0)} \neq 0$ or $\beta^{(1)} \neq 0$.*

$$V_{\text{RR, N}} = \left(\frac{c^{(1)}}{c^{(0)}}\right)^2 \left(\frac{\|\beta^{(1)}\|_\Sigma^2 + \sigma^2}{e(c^{(1)})^2} + \frac{\|\beta^{(0)}\|_\Sigma^2 + \sigma^2}{(1-e)(c^{(0)})^2} \right), \quad (29)$$

$$V_{\text{RR, G}} = \left(\frac{c^{(1)}}{c^{(0)}}\right)^2 \left(\left\| \frac{\beta^{(1)}}{c^{(1)}} \right\|_\Sigma^2 + \left\| \frac{\beta^{(0)}}{c^{(0)}} \right\|_\Sigma^2 - 2 \frac{\langle \beta^{(0)}, \beta^{(1)} \rangle_\Sigma}{c^{(0)}c^{(1)}} \right), \quad (30)$$

and

$$V_{\text{RR, N}} - V_{\text{RR, G}} = \left(\frac{c^{(1)}}{c^{(0)}}\right)^2 \left(e(1-e) \left\| \frac{\beta^{(1)}}{ec^{(1)}} + \frac{\beta^{(0)}}{(1-e)c^{(0)}} \right\|_\Sigma^2 + \sigma^2 \left(\frac{1}{e(c^{(1)})^2} + \frac{1}{(1-e)(c^{(0)})^2} \right) \right) \geq 0 \quad (31)$$

Proof. We first show 29:

Given the following linear model:

$$Y_i^{(t)} = c^{(t)} + X_i^\top \beta^{(t)} + \varepsilon_i^{(t)}, \quad \mathbb{E}[\varepsilon_i^{(t)} | X_i] = 0, \quad \text{Var}[\varepsilon_i^{(t)} | X_i] = \sigma^2.$$

We start by verifying outcome positivity:

$$\begin{aligned} \mathbb{E}[Y^{(a)} | X] &= \mathbb{E}[c(a) + X_i \beta(a) + \varepsilon_i(a) | X] \\ &= c(a) + X_i \beta(a) \end{aligned}$$

Since we have that $\|X\|_\Sigma \leq M$, then for a big enough $c(a)$, outcome positivity is verified. First, we compute the expectation of $Y^{(a)}$:

$$\begin{aligned} \mathbb{E}[Y^{(a)}] &= \mathbb{E}[c(a) + X_i \beta(a) + \varepsilon_i(a)] && \text{(by linear model)} \\ &= \mathbb{E}[c(a)] + \mathbb{E}[X_i \beta(a)] + \mathbb{E}[\varepsilon_i(a)] \\ &= c(a) + \mathbb{E}[X_i] \beta(a) + \mathbb{E}[\mathbb{E}[\varepsilon_i(a) | X]] && \text{(by total expectation)} \\ &= c(a) && \text{(by linear model)} \end{aligned}$$

Next, we compute the variance of $Y^{(a)}$:

$$\begin{aligned}
\text{Var}[Y^{(a)}] &= \text{Var}[c(a) + X\beta(a) + \varepsilon(a)] \\
&= \text{Var}[X\beta(a) + \varepsilon(a)] && c(a) \text{ is a constant} \\
&= \text{Var}[X\beta(a)] + \text{Var}[\varepsilon(a)] + 2 \text{Cov}(X\beta(a), \varepsilon(a)) && \text{Bienaymé's identity} \\
&= \|\beta(a)\|_{\Sigma}^2 + \sigma^2, && \text{(by linear model)}
\end{aligned}$$

where the last equality follows from the following calculations.

We calculate the covariance term:

$$\begin{aligned}
\text{Cov}(X\beta(a), \varepsilon(a)) &= \mathbb{E}[X\beta(a)\varepsilon(a)] - \mathbb{E}[X\beta(a)]\mathbb{E}[\varepsilon(a)] \\
&= \mathbb{E}[X\beta(a)\mathbb{E}[\varepsilon(a)|X]] - \mathbb{E}[X\beta(a)]\mathbb{E}[\mathbb{E}[\varepsilon(a)|X]] && \text{(by total expectation)} \\
&= 0, && \mathbb{E}[\varepsilon(a)|X] = 0
\end{aligned}$$

We also compute the variance of the error term $\varepsilon(a)$ using Eve's law:

$$\begin{aligned}
\text{Var}[\varepsilon(a)] &= \mathbb{E}[\text{Var}[\varepsilon(a)|X]] + \text{Var}[\mathbb{E}[\varepsilon(a)|X]] \\
&= \sigma^2,
\end{aligned}$$

since $\text{Var}[\varepsilon(a)|X] = \sigma^2$ and $\mathbb{E}[\varepsilon(a)|X] = 0$.

Now, using the quantities computed above, we can determine $V_{\text{RR},N}$:

$$V_{\text{RR},N} = \tau_{\text{RR}}^2 \left(\frac{\text{Var}(Y^{(1)})}{e\mathbb{E}[Y^{(1)}]^2} + \frac{\text{Var}(Y^{(0)})}{(1-e)\mathbb{E}[Y^{(0)}]^2} \right).$$

Substituting the expressions for $\mathbb{E}[Y^{(a)}]$ and $\text{Var}[Y^{(a)}]$ obtained above, we get:

$$V_{\text{RR},N} = \left(\frac{c^{(1)}}{c^{(0)}} \right)^2 \left(\frac{\|\beta^{(1)}\|_{\Sigma}^2 + \sigma^2}{e(c^{(1)})^2} + \frac{\|\beta^{(0)}\|_{\Sigma}^2 + \sigma^2}{(1-e)(c^{(0)})^2} \right).$$

We now show 30. Given the linear model, we can compute the following quantities:

First, we compute the variance of $\mu_{\Sigma}^*(X)$:

$$\begin{aligned}
\text{Var}[\mu_{\Sigma}^*(X)] &= \text{Var}[\mathbb{E}[Y^{(a)}|X]] \\
&= \text{Var}[\mathbb{E}[c(a)|X] + \mathbb{E}[X_i\beta(a)|X] + \mathbb{E}[\varepsilon_i(a)|X]] \\
&= \text{Var}[c(a) + \mathbb{E}[X_i\beta(a)|X]] \\
&= \text{Var}[\mathbb{E}[X_i\beta(a)|X]] \\
&= \text{Var}[X_i\beta(a)] \\
&= \|\beta(a)\|_{\Sigma}^2.
\end{aligned}$$

Next, we compute the covariance between $\mu_1^*(X)$ and $\mu_0^*(X)$:

$$\begin{aligned}
\text{Cov}(\mu_1^*(X), \mu_0^*(X)) &= \mathbb{E}[\mu_1^*(X)\mu_0^*(X)] - \mathbb{E}[Y^{(0)}]\mathbb{E}[Y^{(1)}] \\
&= \mathbb{E}[(c(1) + X\beta(1))(c(0) + X\beta(0))] - \mathbb{E}[Y^{(0)}]\mathbb{E}[Y^{(1)}] \\
&= \mathbb{E}[c(1)c(0)] + \mathbb{E}[c(1)X\beta(0)] + \mathbb{E}[c(0)X\beta(1)] \\
&\quad + \mathbb{E}[X\beta(0)X\beta(1)] - \mathbb{E}[Y^{(0)}]\mathbb{E}[Y^{(1)}] \\
&= \mathbb{E}[X\beta(0)X\beta(1)] \\
&= \mathbb{E}\left[\sum_j X_j\beta(0)_j \sum_k X_k\beta(1)_k\right] \\
&= \sum_j \sum_k \beta(0)_j\beta(1)_k \mathbb{E}[X_kX_j] \\
&= \langle \beta^{(0)}, \beta^{(1)} \rangle_{\Sigma}.
\end{aligned}$$

Since we know that:

$$V_{\text{RR,G}} = \tau_{\text{RR}}^2 \text{Var} \left(\frac{\mu_1^*(X)}{\mathbb{E}[Y^{(1)}]} - \frac{\mu_0^*(X)}{\mathbb{E}[Y^{(0)}]} \right),$$

using the computations from above, we get that:

$$V_{\text{RR,G}} = \left(\frac{c^{(1)}}{c^{(0)}} \right)^2 \left(\left\| \frac{\beta^{(1)}}{c^{(1)}} \right\|_{\Sigma}^2 + \left\| \frac{\beta^{(0)}}{c^{(0)}} \right\|_{\Sigma}^2 - 2 \frac{\langle \beta^{(0)}, \beta^{(1)} \rangle_{\Sigma}}{c^{(0)}c^{(1)}} \right).$$

We now show 31 We compute the difference of the two variances:

$$\begin{aligned}
V_{\text{RR,N}} - V_{\text{RR,G}} &= \left(\frac{c^{(1)}}{c^{(0)}} \right)^2 \left(\frac{\|\beta^{(1)}\|_{\Sigma}^2 + \sigma^2}{e(c^{(1)})^2} + \frac{\|\beta^{(0)}\|_{\Sigma}^2 + \sigma^2}{(1-e)(c^{(0)})^2} \right) \\
&\quad - \left(\frac{c^{(1)}}{c^{(0)}} \right)^2 \left(\left\| \frac{\beta^{(1)}}{c^{(1)}} \right\|_{\Sigma}^2 + \left\| \frac{\beta^{(0)}}{c^{(0)}} \right\|_{\Sigma}^2 - 2 \frac{\langle \beta^{(0)}, \beta^{(1)} \rangle_{\Sigma}}{c^{(0)}c^{(1)}} \right) \\
&= \left(\frac{c^{(1)}}{c^{(0)}} \right)^2 \left(\left\| \frac{\beta^{(1)}}{c^{(1)}} \right\|_{\Sigma}^2 \frac{1-e}{e} + \left\| \frac{\beta^{(0)}}{c^{(0)}} \right\|_{\Sigma}^2 \frac{e}{1-e} + 2 \frac{\langle \beta^{(0)}, \beta^{(1)} \rangle_{\Sigma}}{c^{(0)}c^{(1)}} \right. \\
&\quad \left. + \sigma^2 \left(\frac{1}{e(c^{(1)})^2} + \frac{1}{(1-e)(c^{(0)})^2} \right) \right) \\
&= \left(\frac{c^{(1)}}{c^{(0)}} \right)^2 \left(e(1-e) \left\| \frac{\beta^{(1)}}{ec^{(1)}} + \frac{\beta^{(0)}}{(1-e)c^{(0)}} \right\|_{\Sigma}^2 + \sigma^2 \left(\frac{1}{e(c^{(1)})^2} + \frac{1}{(1-e)(c^{(0)})^2} \right) \right) \geq 0
\end{aligned}$$

□

A.2 Proofs of observational studies estimators of the risk ratio

A.2.1 Ratio Inverse Propensity Weighting

Proof of Proposition 4.

Asymptotic bias and variance of the oracle Ratio IPW estimator: we remind that the oracle Ratio IPW is defined as:

$$\tau_{\text{RR,IPW}}^* = \frac{\sum_{i=1}^n \frac{T_i Y_i}{e(X_i)}}{\sum_{i=1}^n \frac{(1-T_i) Y_i}{1-e(X_i)}}$$

where the propensity score e is assumed to be known. Let us define $g_1(Z) = \frac{T_i Y_i}{e(X_i)}$ and $g_0(Z) = \frac{(1-T_i)Y_i}{1-e(X_i)}$ with $Z = (X, T, Y)$. We have that both $g_1(Z)$ and $g_0(Z)$ are bounded and therefore square integrable.

$$\frac{m}{1-\eta} \leq g_1(Z) \leq \frac{M}{\eta} \quad \text{and} \quad g_0(Z) \leq \frac{M}{\eta}$$

We also have that $\mathbb{E}[g_0(Z_i)] = \mathbb{E}[Y^{(0)}]$ and $\mathbb{E}[g_1(Z_i)] = \mathbb{E}[Y^{(1)}]$. We can therefore apply Theorem 2 and conclude that:

$$\sqrt{n}(\tau_{\text{RR,IPW}}^* - \tau_{\text{RR}}) \rightarrow \mathcal{N}(0, V_{\text{RR,IPW}})$$

where $V_{\text{RR,IPW}} = \tau_{\text{RR}}^2 \text{Var}\left(\frac{T_i Y_i}{e(X_i)} - \frac{(1-T_i)Y_i}{1-e(X_i)}\right)$. Moreover,

$$\begin{aligned} \text{Var}\left(\frac{TY}{e(X)}\right) &= \mathbb{E}\left[\left(\frac{TY}{e(X)}\right)^2\right] - \mathbb{E}\left[\frac{TY}{e(X)}\right]^2 \\ &= \mathbb{E}\left[\frac{TY^2}{e(X)^2}\right] - \mathbb{E}[Y^{(1)}]^2 \\ &= \mathbb{E}\left[\frac{1}{e(X)^2} \mathbb{E}[TY^{(1)2}|X]\right] - \mathbb{E}[Y^{(1)}]^2 \\ &= \mathbb{E}\left[\frac{1}{e(X)} \mathbb{E}[Y^{(1)2}|X]\right] - \mathbb{E}[Y^{(1)}]^2 \\ &= \mathbb{E}\left[\frac{1}{e(X)} \mathbb{E}[Y^{(1)2}|X]\right] - \mathbb{E}[Y^{(1)}]^2 \\ &= \mathbb{E}\left[\frac{Y^{(1)2}}{e(X)}\right] - \mathbb{E}[Y^{(1)}]^2 \end{aligned}$$

Similarly:

$$\text{Var}\left(\frac{(1-T)Y}{1-e(X)}\right) = \mathbb{E}\left[\frac{Y^{(0)2}}{1-e(X)}\right] - \mathbb{E}[Y^{(0)}]^2$$

And for the covariance:

$$\begin{aligned} \text{Cov}\left(\frac{TY}{e(X)}, \frac{(1-T)Y}{1-e(X)}\right) &= \mathbb{E}\left[\left(\frac{TY}{e(X)} - \mathbb{E}[Y^{(1)}]\right)\left(\frac{(1-T)Y}{1-e(X)} - \mathbb{E}[Y^{(0)}]\right)\right] \\ &= \mathbb{E}\left[\frac{TY}{e(X)} \frac{(1-T)Y}{1-e(X)}\right] - \mathbb{E}[Y^{(1)}] \mathbb{E}\left[\frac{(1-T)Y}{1-e(X)}\right] \\ &\quad - \mathbb{E}[Y^{(0)}] \mathbb{E}\left[\frac{TY}{e(X)}\right] + \mathbb{E}[Y^{(1)}] \mathbb{E}[Y^{(0)}] \\ &= -\mathbb{E}[Y^{(1)}] \mathbb{E}[Y^{(0)}] \end{aligned}$$

Therefore we get that:

$$V_{\text{RR,IPW}} = \tau_{\text{RR}}^2 \left(\frac{\mathbb{E}\left[\frac{Y^{(1)2}}{e(X)}\right]}{\mathbb{E}[Y^{(1)}]^2} + \frac{\mathbb{E}\left[\frac{Y^{(0)2}}{1-e(X)}\right]}{\mathbb{E}[Y^{(0)}]^2} \right)$$

Finite sample bias and variance of the oracle Ratio IPW estimator: let $T_1(\mathbf{Z}) = \frac{1}{n} \sum_{i=1}^n \frac{T_i Y_i}{e(X_i)}$ and $T_0(\mathbf{Z}) = \frac{1}{n} \sum_{i=1}^n \frac{(1-T_i)Y_i}{1-e(X_i)}$ where $\mathbf{Z} = (Z_1, \dots, Z_n)$. We first show that $\text{Var}(T_1(\mathbf{Z})) = O_p\left(\frac{1}{n}\right)$ and $\text{Var}(T_0(\mathbf{Z})) = O_p\left(\frac{1}{n}\right)$:

$$\begin{aligned}
\text{Var}(T_1(\mathbf{Z})) &= \frac{1}{n^2} \text{Var} \left(\sum_{i=1}^n \frac{T_i Y_i}{e(X_i)} \right) \\
&= \frac{1}{n^2} \sum_{i=1}^n \text{Var} \left(\frac{T_i Y_i}{e(X_i)} \right) && \text{(by i.i.d.)} \\
&= \frac{1}{n} \left(\mathbb{E} \left[\left(\frac{T_i Y_i}{e(X_i)} \right)^2 \right] - \mathbb{E} \left[\frac{T_i Y_i}{e(X_i)} \right]^2 \right) && \text{(by law of total expectation)} \\
&= \frac{\mathbb{E} \left[\frac{(Y^{(1)})^2}{e(X_i)} \right] - \mathbb{E} [Y^{(1)}]^2}{n} \\
&= O_p \left(\frac{1}{n} \right)
\end{aligned}$$

Similarly,

$$\text{Var}(T_0(\mathbf{Z})) = O_p \left(\frac{1}{n} \right)$$

Therefore, we have shown that $T_1(\mathbf{Z})$ and $T_0(\mathbf{Z})$ are unbiased estimators of $\mathbb{E} [Y^{(1)}]$ and $\mathbb{E} [Y^{(0)}] > 0$, respectively. We also have that $T_0(\mathbf{Z})$ and $T_1(\mathbf{Z})$ are bounded:

$$\frac{m}{1-\eta} \leq T_0(\mathbf{Z}) \leq \frac{M}{\eta} \quad \text{and} \quad T_1(\mathbf{Z}) \leq \frac{M}{\eta}$$

Applying Theorem 2, we obtain:

$$|\mathbb{E} [\hat{\tau}_{\text{RR, HT, n}}] - \tau_{\text{RR}}| \leq \frac{2M^3(1-\eta)^3}{nm^3\eta^3}$$

and

$$|\text{Var}(\hat{\tau}_{\text{RR, HT, n}}) - V_{\text{RR, HT}}| \leq \frac{4M^4(1-\eta)^6}{nm^6\eta^4}$$

□

A.2.2 Ratio Augmented Inverse Propensity Weighting

Proof of Proposition 5.

Asymptotic bias and variance of the crossfitted Ratio AIPW estimator: Recall that we want to analyze $\sqrt{n}(\tilde{\tau}_{\text{RR, AIPW}} - \tau_{\text{RR}})$. Letting

$$\tau_{\text{RR, AIPW}}^* = \frac{\sum_{i=1}^n \mu_1(X_i) + \frac{T_i(Y_i - \mu_1(X_i))}{e(X_i)}}{\sum_{i=1}^n \mu_0(X_i) + \frac{(1-T_i)(Y_i - \mu_0(X_i))}{1-e(X_i)}} := \frac{\tau_{\text{RR, AIPW, 1}}^*}{\tau_{\text{RR, AIPW, 0}}^*}, \quad (32)$$

be the oracle version of $\tilde{\tau}_{\text{RR, AIPW}}$ where the propensity score and both response surfaces are assumed to be known, we can rewrite

$$\tilde{\tau}_{\text{RR, AIPW}} - \tau_{\text{RR}} = \tilde{\tau}_{\text{RR, AIPW}} - \tau_{\text{RR, AIPW}}^* + \tau_{\text{RR, AIPW}}^* - \tau_{\text{RR}}. \quad (33)$$

Regarding the first term in (33), we have

$$\begin{aligned}
\left| \tilde{\tau}_{\text{RR, AIPW}} - \tau_{\text{RR, AIPW}}^* \right| &= \left| \frac{\tilde{\tau}_{\text{RR, AIPW, 1}}}{\tilde{\tau}_{\text{RR, AIPW, 0}}} - \frac{\tau_{\text{RR, AIPW, 1}}^*}{\tau_{\text{RR, AIPW, 0}}^*} \right| \\
&= \left| \left((\tilde{\tau}_{\text{RR, AIPW, 0}})^{-1} - (\tau_{\text{RR, AIPW, 0}}^*)^{-1} \right) \tilde{\tau}_{\text{RR, AIPW, 1}} \right. \\
&\quad \left. + (\tau_{\text{RR, AIPW, 0}}^*)^{-1} (\tilde{\tau}_{\text{RR, AIPW, 1}} - \tau_{\text{RR, AIPW, 1}}^*) \right| \\
&\leq \left| \left((\tilde{\tau}_{\text{RR, AIPW, 0}})^{-1} - (\tau_{\text{RR, AIPW, 0}}^*)^{-1} \right) \tilde{\tau}_{\text{RR, AIPW, 1}} \right| \\
&\quad + \left| (\tau_{\text{RR, AIPW, 0}}^*)^{-1} (\tilde{\tau}_{\text{RR, AIPW, 1}} - \tau_{\text{RR, AIPW, 1}}^*) \right|
\end{aligned}$$

We now show that:

$$|\tilde{\tau}_{\text{RR,AIPW},1} - \tau_{\text{RR,AIPW},1}^*| = o_p\left(\frac{1}{\sqrt{n}}\right) \quad \text{and} \quad |\tilde{\tau}_{\text{RR,AIPW},0} - \tau_{\text{RR,AIPW},0}^*| = o_p\left(\frac{1}{\sqrt{n}}\right)$$

One can show that:

$$\begin{aligned} \sqrt{n} |\tilde{\tau}_{\text{RR,AIPW},1} - \tau_{\text{RR,AIPW},1}^*| &= \frac{1}{\sqrt{n}} \sum_{i \in \mathcal{I}_k} \left(\hat{\mu}_1^{\mathcal{I}-k}(X_i) + T_i \frac{Y_i - \hat{\mu}_1^{\mathcal{I}-k}(X_i)}{\hat{e}(X_i)} - \mu_1(X_i) - T_i \frac{Y_i - \mu_1(X_i)}{e(X_i)} \right) \\ \text{Further denoted } A_n^k &= \frac{1}{\sqrt{n}} \sum_{i \in \mathcal{I}_k} \left(\left(\hat{\mu}_1^{\mathcal{I}-k}(X_i) - \mu_1(X_i) \right) \left(1 - \frac{T_i}{e(X_i)} \right) \right) \\ \text{Further denoted } B_n^k &+ \frac{1}{\sqrt{n}} \sum_{i \in \mathcal{I}_k} T_i \left((Y_i - \mu_1(X_i)) \left(\frac{1}{\hat{e}(X_i)} - \frac{1}{e(X_i)} \right) \right) \\ \text{Further denoted } C_n^k &- \frac{1}{\sqrt{n}} \sum_{i \in \mathcal{I}_k} T_i \left(\left(\hat{\mu}_1^{\mathcal{I}-k}(X_i) - \mu_1(X_i) \right) \left(\frac{1}{\hat{e}(X_i)} - \frac{1}{e(X_i)} \right) \right). \end{aligned}$$

The first two terms can be demonstrated to converge toward 0 in probability. To do this, one can show that each of these two first terms converge in L^2 norm.

First, one can show that the expectation of $\frac{A_n^k}{\sqrt{n}}$ is null,

$$\begin{aligned} \mathbb{E} \left[\frac{A_n^k}{\sqrt{n}} \right] &= \frac{1}{n} \sum_{i \in \mathcal{I}_k} \mathbb{E} \left[\left(\hat{\mu}_1^{\mathcal{I}-k}(X_i) - \mu_1(X_i) \right) \left(1 - \frac{T_i}{e(X_i)} \right) \right] \\ &= \frac{|\mathcal{I}_k|}{n} \mathbb{E} \left[\left(\hat{\mu}_1^{\mathcal{I}-k}(X) - \mu_1(X) \right) \left(1 - \frac{T}{e(X)} \right) \right] \quad \text{i.i.d.} \\ &= \frac{|\mathcal{I}_k|}{n} \mathbb{E} \left[\mathbb{E} \left[\left(\hat{\mu}_1^{\mathcal{I}-k}(X) - \mu_1(X) \right) \left(1 - \frac{T}{e(X)} \right) \mid X \right] \right] \\ &= \frac{|\mathcal{I}_k|}{n} \mathbb{E} \left[\mathbb{E} \left[\left(\hat{\mu}_1^{\mathcal{I}-k}(X) - \mu_1(X) \right) \left(1 - \frac{\mathbb{E}[T \mid X]}{e(X)} \right) \right] \right] \\ &= \frac{|\mathcal{I}_k|}{n} \mathbb{E} \left[\mathbb{E} \left[\left(\hat{\mu}_1^{\mathcal{I}-k}(X) - \mu_1(X) \right) \left(1 - \frac{e(X)}{e(X)} \right) \right] \right] \\ &= 0. \end{aligned}$$

This will be helpful in the next series of derivations.

Now, consider the expectation of the square of $\frac{A_n^k}{\sqrt{n}}$,

$$\begin{aligned}
\mathbb{E} \left[\left(\frac{A_n^k}{\sqrt{n}} \right)^2 \right] &= \text{Var} \left[\frac{1}{n} \sum_{i \in \mathcal{I}_k} \left(\left(\hat{\mu}_1^{\mathcal{I}-k}(X_i) - \mu_1(X_i) \right) \left(1 - \frac{T_i}{e(X_i)} \right) \right) \right] \\
&= \frac{1}{n^2} \text{Var} \left[\sum_{i \in \mathcal{I}_k} \left(\left(\hat{\mu}_1^{\mathcal{I}-k}(X_i) - \mu_1(X_i) \right) \left(1 - \frac{T_i}{e(X_i)} \right) \right) \right] \\
&= \frac{1}{n^2} \sum_{i \in \mathcal{I}_k} \text{Var} \left[\left(\hat{\mu}_1^{\mathcal{I}-k}(X_i) - \mu_1(X_i) \right) \left(1 - \frac{T_i}{e(X_i)} \right) \right] \quad \text{iid} \\
&= \frac{|\mathcal{I}_k|}{n^2} \mathbb{E} \left[\left(\left(\hat{\mu}_1^{\mathcal{I}-k}(X_i) - \mu_1(X_i) \right) \left(1 - \frac{T_i}{e(X_i)} \right) \right)^2 \right] \\
&= \frac{|\mathcal{I}_k|}{n^2} \mathbb{E} \left[\mathbb{E} \left[\left(\left(\hat{\mu}_1^{\mathcal{I}-k}(X_i) - \mu_1(X_i) \right) \left(1 - \frac{T_i}{e(X_i)} \right) \right)^2 \mid X \right] \right] \\
&= \frac{|\mathcal{I}_k|}{n^2} \mathbb{E} \left[\left(\hat{\mu}_1^{\mathcal{I}-k}(X_i) - \mu_1(X_i) \right)^2 \mathbb{E} \left[\left(1 - \frac{T_i}{e(X_i)} \right)^2 \mid X \right] \right] \\
&= \frac{|\mathcal{I}_k|}{n^2} \mathbb{E} \left[\left(\hat{\mu}_1^{\mathcal{I}-k}(X_i) - \mu_1(X_i) \right)^2 \frac{1}{e(X_i)^2} \mathbb{E} \left[(e(X_i) - T_i)^2 \mid X \right] \right] \\
&= \frac{|\mathcal{I}_k|}{n^2} \mathbb{E} \left[\left(\hat{\mu}_1^{\mathcal{I}-k}(X_i) - \mu_1(X_i) \right)^2 \frac{e(X_i)(1 - e(X_i))}{e(X_i)^2} \right] \\
&= \frac{|\mathcal{I}_k|}{n^2} \mathbb{E} \left[\left(\hat{\mu}_1^{\mathcal{I}-k}(X_i) - \mu_1(X_i) \right)^2 \left(\frac{1}{e(X_i)} - 1 \right) \right] \\
&\leq \frac{|\mathcal{I}_k|}{\eta n^2} \mathbb{E} \left[\left(\hat{\mu}_1^{\mathcal{I}-k}(X_i) - \mu_1(X_i) \right)^2 \right] \quad \text{Overlap} \\
&\leq \frac{|\mathcal{I}_k|}{n^2} o_{\mathbb{P}}(1)
\end{aligned}$$

Therefore, because convergence on L^2 -norm provides convergence in probability (Chebyshev inequality), we have for $k \in \{1, K\}$:

$$\sqrt{n} \frac{1}{n} \sum_{i \in \mathcal{I}_k} \left(\left(\hat{\mu}_1^{\mathcal{I}-k}(X_i) - \mu_1(X_i) \right) \left(1 - \frac{T_i}{e(X_i)} \right) \right) \xrightarrow{p} 0$$

The second term can also be controlled using similar arguments. Before detailed derivation, note that due to the uniform convergence of $\hat{e}(\cdot)$ and the overlap assumption, there exist M such that for all $n > M$, and for all X_i ,

$$\frac{\eta}{2} \leq \hat{e}(X_i) \leq 1 - \frac{\eta}{2}.$$

Therefore, there exist M such that for all $n > M$, and for all X_i ,

$$\begin{aligned}
\frac{1}{\hat{e}(X)} - \frac{1}{e(X)} &= \frac{e(X) - \hat{e}(X)}{\hat{e}(X)e(X)} \\
&\leq 2 \frac{e(X) - \hat{e}(X)}{\eta^2}.
\end{aligned}$$

Derivations are very close to the ones for the first term, noting that,

$$\mathbb{E} \left[\mathbb{E} \left[\frac{1}{n} \sum_{i \in \mathcal{I}_k} T_i \left((Y_i - \mu_1(X_i)) \left(\frac{1}{\hat{e}^{\mathcal{I}-k}(X_i)} - \frac{1}{e(X_i)} \right) \right) \mid X_i \right] \right] = 0,$$

so that,

$$\begin{aligned}
\mathbb{E} \left[\left(\frac{B_n^k}{\sqrt{n}} \right)^2 \right] &= \text{Var} \left[\frac{1}{n} \sum_{i \in \mathcal{I}_k} T_i (Y_i - \mu_1(X_i)) \left(\frac{1}{\hat{e}^{\mathcal{I}-k}(X_i)} - \frac{1}{e(X_i)} \right) \right] \\
&= \frac{1}{n^2} \sum_{i \in \mathcal{I}_k} \text{Var} \left[T_i (Y_i - \mu_1(X_i)) \left(\frac{1}{\hat{e}^{\mathcal{I}-k}(X_i)} - \frac{1}{e(X_i)} \right) \right] \quad \text{iid} \\
&= \frac{|\mathcal{I}_k|}{n^2} \mathbb{E} \left[T (Y - \mu_1(X))^2 \left(\frac{1}{\hat{e}^{\mathcal{I}-k}(X)} - \frac{1}{e(X)} \right)^2 \right] \\
&\leq \frac{4|\mathcal{I}_k|}{\eta^4 n^2} \mathbb{E} \left[T (Y - \mu_1(X))^2 (\hat{e}^{\mathcal{I}-k}(X) - e(X))^2 \right] \\
&\leq \frac{4|\mathcal{I}_k|}{\eta^4 n^2} \mathbb{E} \left[(Y - \mu_1(X))^2 (\hat{e}^{\mathcal{I}-k}(X) - e(X))^2 \right] \quad \text{Since } T \leq 1 \\
&\leq \frac{4|\mathcal{I}_k|}{\eta^4 n^2} \mathbb{E} \left[\mathbb{E} \left[(Y - \mu_1(X))^2 (\hat{e}^{\mathcal{I}-k}(X) - e(X))^2 \mid X \right] \right] \\
&\leq \frac{4|\mathcal{I}_k|}{\eta^4 n^2} \mathbb{E} \left[\mathbb{E} \left[(Y - \mu_1(X))^2 \mid X \right] (\hat{e}^{\mathcal{I}-k}(X) - e(X))^2 \right] \\
&\leq \frac{4|\mathcal{I}_k|}{\eta^4 n^2} \mathbb{E} \left[\text{Var} [Y \mid X] (\hat{e}^{\mathcal{I}-k}(X) - e(X))^2 \right] \\
&\leq \frac{4 \text{Var} [Y \mid X] |\mathcal{I}_k|}{\eta^4 n^2} \mathbb{E} \left[(\hat{e}^{\mathcal{I}-k}(X) - e(X))^2 \right] \\
&\leq \frac{4 \text{Var} [Y \mid X] |\mathcal{I}_k|}{\eta^4 n^2} o_{\mathbb{P}}(1)
\end{aligned}$$

Therefore, for $k \in \{1, K\}$:

$$\sqrt{n} \frac{1}{n} \sum_{i \in \mathcal{I}_k} T_i (Y_i - \mu_1(X_i)) \left(\frac{1}{\hat{e}^{\mathcal{I}-k}(X_i)} - \frac{1}{e(X_i)} \right) \xrightarrow{p} 0.$$

For the last term, the approach is different and will involve another assumption on the product of residuals,

$$\begin{aligned}
C_n^k &= \sqrt{n} \frac{1}{n} \sum_{i \in \mathcal{I}_k} T_i \left(\hat{\mu}_1^{\mathcal{I}-k}(X_i) - \mu_1(X_i) \right) \left(\frac{1}{\hat{e}^{\mathcal{I}-k}(X_i)} - \frac{1}{e(X_i)} \right) \\
&\leq \sqrt{n} \sqrt{\frac{1}{n} \sum_{i \in \mathcal{I}_k} T_i \left(\hat{\mu}_1^{\mathcal{I}-k}(X_i) - \mu_1(X_i) \right)^2} \sqrt{\frac{1}{n} \sum_{i \in \mathcal{I}_k} \left(\frac{1}{\hat{e}^{\mathcal{I}-k}(X_i)} - \frac{1}{e(X_i)} \right)^2} \quad \text{C.S.} \\
&\leq \frac{\sqrt{n}}{\eta} \sqrt{\frac{1}{n} \sum_{i \in \mathcal{I}_k} T_i \left(\hat{\mu}_1^{\mathcal{I}-k}(X_i) - \mu_1(X_i) \right)^2} \sqrt{\frac{1}{n} \sum_{i \in \mathcal{I}_k} (e(X_i) - \hat{e}^{\mathcal{I}-k}(X_i))^2} \quad \text{Overlap} \\
&= \frac{\sqrt{n}}{\eta} o_{\mathbb{P}}\left(\frac{1}{\sqrt{n}}\right) \quad \text{Assumption} \\
&= \frac{1}{\eta} o_{\mathbb{P}}(1)
\end{aligned}$$

Each term A_n^k , B_n^k , and C_n^k has been shown to be bounded by a term in $o_{\mathbb{P}}(1)$. The remainder term with function $\hat{m}_0^k(Y, A, X) - m_0(Y, A, X)$ can also be controlled with the same derivation, except

using the uniform convergence of $\hat{\mu}_0^{\mathcal{I}-k}(\cdot)$. Since we have:

$$\sqrt{n}(\hat{\tau}_{\text{AIPW}} - \hat{\tau}_{\text{AIPW}}^*) = \sum_{k=1}^K A_n^k + B_n^k + C_n^k$$

Therefore,

$$\sqrt{n}(\tilde{\tau}_{\text{RR,AIPW},1} - \tau_{\text{RR,AIPW},1}^*) \xrightarrow{p} 0$$

. And similarly:

$$\sqrt{n}(\tilde{\tau}_{\text{RR,AIPW},0} - \tau_{\text{RR,AIPW},0}^*) \xrightarrow{p} 0$$

Therefore, we have that:

$$\begin{aligned} |\tilde{\tau}_{\text{RR,AIPW}} - \tau_{\text{RR,AIPW}}^*| &\leq \left| \left((\tilde{\tau}_{\text{RR,AIPW},0})^{-1} - (\tau_{\text{RR,AIPW},0}^*)^{-1} \right) \tilde{\tau}_{\text{RR,AIPW},1} \right| \\ &\quad + \left| (\tau_{\text{RR,AIPW},0}^*)^{-1} (\tilde{\tau}_{\text{RR,AIPW},1} - \tau_{\text{RR,AIPW},1}^*) \right| \\ &\leq \frac{|\tilde{\tau}_{\text{RR,AIPW}}|}{\tau_{\text{RR,AIPW},0}^*} o_p\left(\frac{1}{\sqrt{n}}\right) + \left| (\tau_{\text{RR,AIPW},0}^*)^{-1} \right| o_p\left(\frac{1}{\sqrt{n}}\right) \end{aligned}$$

Consequently, in probability, as n tends to infinity,

$$\sqrt{n} |\tilde{\tau}_{\text{RR,AIPW}} - \tau_{\text{RR,AIPW}}^*| \rightarrow 0. \quad (34)$$

Regarding the second term in (33), we can use Theorem 1 with $g_1(Z) = \mu_1(X_i) + \frac{T_i(Y_i - \mu_1(X_i))}{e(X_i)}$ and $g_1(Z) = \mu_1(X_i) + \frac{T_i(Y_i - \mu_1(X_i))}{e(X_i)}$ where $Z = (T, X, Y)$. Hence, we have that $g_1(Z)$ is square integrable:

$$\begin{aligned} \mathbb{E} [g_1(Z_i)^2] &= \mathbb{E} \left[\left(\mu_1(X_i) + \frac{T_i(Y_i - \mu_1(X_i))}{e(X_i)} \right)^2 \right] \\ &= \mathbb{E} [(\mu_1(X_i))^2] + \mathbb{E} \left[\left(\frac{T_i(Y_i - \mu_1(X_i))}{e(X_i)} \right)^2 \right] \\ &\quad + 2\mathbb{E} \left[\mu_1(X_i) \frac{T_i(Y_i - \mu_1(X_i))}{e(X_i)} \right] \end{aligned}$$

We have that $\mathbb{E}[(\mu_1(X_i))^2] = \text{Var}(Y^{(1)}) + \mathbb{E}[Y^{(1)}]^2$ and therefore is finite. Using Consistency, Unconfoundedness, and definition of $\mu_1(X) = \mathbb{E}[Y | X, T = 1]$,

$$\begin{aligned}
\mathbb{E} \left[\left(T \frac{Y - \mu_1(X)}{e(X)} \right)^2 \right] &= \mathbb{E} \left[\left(T \frac{Y^{(1)} - \mu_1(X)}{e(X)} \right)^2 \right] && \text{Consistency} \\
&= \mathbb{E} \left[\mathbb{E} \left[\left(T \frac{Y^{(1)} - \mu_1(X)}{e(X)} \right)^2 \mid X \right] \right] && \text{Total expectation} \\
&= \mathbb{E} \left[\mathbb{E} \left[T \left(\frac{Y^{(1)} - \mu_1(X)}{e(X)} \right)^2 \mid X \right] \right] && \text{T is binary} \\
&= \mathbb{E} \left[\mathbb{E} \left[\mathbf{1}_{\{T=1\}} \left(\frac{Y^{(1)} - \mu_1(X)}{e(X)} \right)^2 \mid X \right] \right] && \text{T written as an indicator} \\
&= \mathbb{E} \left[\frac{1}{e(X)^2} \mathbb{E} \left[\mathbf{1}_{\{T=1\}} \left(Y^{(1)} - \mu_1(X) \right)^2 \mid X \right] \right] && e(X) \text{ is a function of } X \\
&= \mathbb{E} \left[\frac{\left(Y^{(1)} - \mu_1(X) \right)^2}{e(X)^2} \mathbb{E} \left[\mathbf{1}_{\{T=1\}} \mid X \right] \right] && \text{Unconf. \& } \mu_1(\cdot) \text{ is func. of } X \\
&= \mathbb{E} \left[\frac{\left(Y^{(1)} - \mu_1(X) \right)^2}{e(X)^2} e(X) \right] && \text{Definition of } e(X) \\
&= \mathbb{E} \left[\frac{\left(Y^{(1)} - \mu_1(X) \right)^2}{e(X)} \right] \\
&\leq \frac{\text{Var}(\mu_1(X))}{\eta}
\end{aligned}$$

And for the last term we have that:

$$\mathbb{E} \left[\mu_1(X_i) \frac{T_i(Y_i - \mu_1(X_i))}{e(X_i)} \right] = 0$$

Similarly we can show that $g_0(Z)$ is square integrable.

We also have that $\mathbb{E}[g_0(Z_i)] = \mathbb{E}[Y^{(0)}]$ and $\mathbb{E}[g_1(Z_i)] = \mathbb{E}[Y^{(1)}]$. We can therefore apply Theorem 1 and conclude that:

$$\sqrt{n}(\tau_{\text{RR,AIPW}}^* - \tau_{\text{RR}}) \rightarrow \mathcal{N}(0, V_{\text{RR,AIPW}}) \quad (35)$$

where $V_{\text{RR,AIPW}} = \tau_{\text{RR}}^2 \text{Var} \left(\frac{g_1(Z)}{\mathbb{E}[Y^{(1)}]} - \frac{g_0(Z)}{\mathbb{E}[Y^{(0)}]} \right)$. Using Bienaymé's identity we get:

$$\begin{aligned}
&\text{Var} \left(\frac{g_1(Z)}{\mathbb{E}[Y^{(1)}]} - \frac{g_0(Z)}{\mathbb{E}[Y^{(0)}]} \right) \\
&= \text{Var} \left(\frac{\mu_1(X)}{\mathbb{E}[Y^{(1)}]} - \frac{\mu_0(X)}{\mathbb{E}[Y^{(0)}]} \right) + \text{Var} \left(\frac{T(Y - \mu_1(X))}{\mathbb{E}[Y^{(1)}] e(X)} - \frac{(1-T)(Y - \mu_0(X))}{\mathbb{E}[Y^{(0)}] (1 - e(X))} \right) \\
&+ 2 \text{Cov} \left(\frac{\mu_1(X)}{\mathbb{E}[Y^{(1)}]} - \frac{\mu_0(X)}{\mathbb{E}[Y^{(0)}]}, \frac{T(Y - \mu_1(X))}{\mathbb{E}[Y^{(1)}] e(X)} - \frac{(1-T)(Y - \mu_0(X))}{\mathbb{E}[Y^{(0)}] (1 - e(X))} \right)
\end{aligned}$$

For the second term:

$$\begin{aligned} & \text{Var} \left(\frac{T(Y - \mu_1(X))}{\mathbb{E}[Y^{(1)}] e(X)} - \frac{(1-T)(Y - \mu_0(X))}{\mathbb{E}[Y^{(0)}] (1 - e(X))} \right) \\ &= \text{Var} \left(\frac{T(Y - \mu_1(X))}{\mathbb{E}[Y^{(1)}] e(X)} \right) + \text{Var} \left(\frac{(1-T)(Y - \mu_0(X))}{\mathbb{E}[Y^{(0)}] (1 - e(X))} \right) \\ & \quad - 2 \text{Cov} \left(\frac{T(Y - \mu_1(X))}{\mathbb{E}[Y^{(1)}] e(X)}, \frac{(1-T)(Y - \mu_0(X))}{\mathbb{E}[Y^{(0)}] (1 - e(X))} \right) \end{aligned}$$

And since we have that:

$$\text{Var} \left(\frac{T(Y - \mu_1(X))}{\mathbb{E}[Y^{(1)}] e(X)} \right) = \mathbb{E} \left[\left(\frac{T(Y - \mu_1(X))}{\mathbb{E}[Y^{(1)}] e(X)} \right)^2 \right] - \mathbb{E} \left[\frac{T(Y - \mu_1(X))}{\mathbb{E}[Y^{(1)}] e(X)} \right]^2$$

For the first term:

$$\begin{aligned} & \mathbb{E} \left[\left(T \frac{Y - \mu_1(X)}{e(X) \mathbb{E}[Y^{(1)}]} \right)^2 \right] \\ &= \mathbb{E} \left[\left(T \frac{Y^{(1)} - \mu_1(X)}{e(X) \mathbb{E}[Y^{(1)}]} \right)^2 \right] && \text{Consistency} \\ &= \mathbb{E} \left[\mathbb{E} \left[\left(T \frac{Y^{(1)} - \mu_1(X)}{e(X) \mathbb{E}[Y^{(1)}]} \right)^2 \mid X \right] \right] && \text{Total expectation} \\ &= \mathbb{E} \left[\mathbb{E} \left[T \left(\frac{Y^{(1)} - \mu_1(X)}{e(X) \mathbb{E}[Y^{(1)}]} \right)^2 \mid X \right] \right] && \text{T is binary} \\ &= \mathbb{E} \left[\mathbb{E} \left[\mathbf{1}_{\{T=1\}} \left(\frac{Y^{(1)} - \mu_1(X)}{e(X) \mathbb{E}[Y^{(1)}]} \right)^2 \mid X \right] \right] && \text{T written as an indicator} \\ &= \mathbb{E} \left[\frac{1}{e(X)^2 \mathbb{E}[Y^{(1)}]^2} \mathbb{E} \left[\mathbf{1}_{\{T=1\}} \left(Y^{(1)} - \mu_1(X) \right)^2 \mid X \right] \right] && e(X) \text{ is a function of } X \\ &= \mathbb{E} \left[\frac{\text{Var}(Y^{(1)}|X)}{e(X)^2 \mathbb{E}[Y^{(1)}]^2} \mathbb{E}[\mathbf{1}_{\{T=1\}} \mid X] \right] && \text{Uncounf. \& } \mu_1(\cdot) \text{ is func. of } X \\ &= \mathbb{E} \left[\frac{\text{Var}(Y^{(1)}|X)}{e(X)^2 \mathbb{E}[Y^{(1)}]^2} e(X) \right] && \text{Definition of } e(X) \\ &= \mathbb{E} \left[\frac{\text{Var}(Y^{(1)}|X)}{e(X) \mathbb{E}[Y^{(1)}]^2} \right], \end{aligned}$$

For the second term:

$$\begin{aligned}
& \mathbb{E} \left[\frac{T(Y - \mu_1(X))}{\mathbb{E}[Y^{(1)}] e(X)} \right] \\
&= \mathbb{E} \left[\frac{T(Y^{(1)} - \mu_1(X))}{\mathbb{E}[Y^{(1)}] e(X)} \right] && \text{Consistency} \\
&= \mathbb{E} \left[\mathbb{E} \left[T \frac{Y^{(1)} - \mu_1(X)}{e(X) \mathbb{E}[Y^{(1)}]} \mid X \right] \right] && \text{Total expectation} \\
&= \mathbb{E} \left[\frac{1}{e(X) \mathbb{E}[Y^{(1)}]} \mathbb{E} \left[T(Y^{(1)} - \mu_1(X)) \mid X \right] \right] && e(X) \text{ is a function of } X \\
&= \mathbb{E} \left[\frac{e(X)}{e(X) \mathbb{E}[Y^{(1)}]} (\mu_1(X) - \mu_1(X)) \right] && \text{Uncounf. \& } \mu_1(\cdot) \text{ is func. of } X \\
&= 0
\end{aligned}$$

Therefore:

$$\text{Var} \left(\frac{T(Y - \mu_1(X))}{\mathbb{E}[Y^{(1)}] e(X)} \right) = \mathbb{E} \left[\frac{\text{Var}(Y^{(1)}|X)}{e(X) \mathbb{E}[Y^{(1)}]^2} \right]$$

Similarly:

$$\text{Var} \left(\frac{(1-T)(Y - \mu_0(X))}{\mathbb{E}[Y^{(0)}] (1-e(X))} \right) = \mathbb{E} \left[\frac{\text{Var}(Y^{(0)}|X)}{(1-e(X)) \mathbb{E}[Y^{(0)}]^2} \right]$$

And since we also have that:

$$\begin{aligned}
\text{Cov} \left(\frac{T(Y - \mu_1(X))}{\mathbb{E}[Y^{(1)}] e(X)}, \frac{(1-T)(Y - \mu_0(X))}{\mathbb{E}[Y^{(0)}] (1-e(X))} \right) &= \mathbb{E} \left[\frac{T(Y - \mu_1(X)) (1-T)(Y - \mu_0(X))}{\mathbb{E}[Y^{(1)}] e(X) \mathbb{E}[Y^{(0)}] (1-e(X))} \right] \\
&\quad - \mathbb{E} \left[\frac{T(Y - \mu_1(X))}{\mathbb{E}[Y^{(1)}] e(X)} \right] \mathbb{E} \left[\frac{(1-T)(Y - \mu_0(X))}{\mathbb{E}[Y^{(0)}] (1-e(X))} \right] \\
&= 0
\end{aligned}$$

We get that:

$$\text{Var} \left(\frac{T(Y - \mu_1(X))}{\mathbb{E}[Y^{(1)}] e(X)} - \frac{(1-T)(Y - \mu_0(X))}{\mathbb{E}[Y^{(0)}] (1-e(X))} \right) = \mathbb{E} \left[\frac{\text{Var}(Y^{(1)}|X)}{e(X) \mathbb{E}[Y^{(1)}]^2} \right] + \mathbb{E} \left[\frac{\text{Var}(Y^{(0)}|X)}{(1-e(X)) \mathbb{E}[Y^{(0)}]^2} \right]$$

Moreover we also have that:

$$\text{Cov} \left(\frac{\mu_1(X)}{\mathbb{E}[Y^{(1)}]} - \frac{\mu_0(X)}{\mathbb{E}[Y^{(0)}]}, \frac{T(Y - \mu_1(X))}{\mathbb{E}[Y^{(1)}] e(X)} - \frac{(1-T)(Y - \mu_0(X))}{\mathbb{E}[Y^{(0)}] (1-e(X))} \right) = 0$$

Hence, we have that:

$$\begin{aligned}
\text{Cov} \left(\mu_1(X), \frac{T(Y - \mu_1(X))}{e(X)} \right) &= \mathbb{E} \left[\mu_1(X) \frac{T(Y - \mu_1(X))}{e(X)} \right] - \mathbb{E}[\mu_1(X)] \mathbb{E} \left[\frac{T(Y - \mu_1(X))}{e(X)} \right] \\
&= \mathbb{E} \left[\mu_1(X) \frac{T(Y - \mu_1(X))}{e(X)} \right] \\
&= \mathbb{E} \left[\frac{\mu_1(X)}{e(X)} \mathbb{E} \left[T(Y^{(1)} - \mu_1(X)) \mid X \right] \right] \\
&= \mathbb{E} \left[\frac{\mu_1(X)}{e(X)} \mathbb{E} \left[\frac{T(Y^{(1)} - \mu_1(X))}{e(X)} \mid X \right] \right] \\
&= \mathbb{E} \left[\frac{\mu_1(X) e(X)}{e(X)} (\mu_1(X) - \mu_1(X)) \right] \\
&= 0
\end{aligned}$$

We similarly show that all other terms are null:

$$\begin{aligned} \text{Cov}\left(\mu_0(X), \frac{T(Y - \mu_1(X))}{e(X)}\right) &= 0 & \text{Cov}\left(\mu_1(X), \frac{(1-T)(Y - \mu_0(X))}{1-e(X)}\right) &= 0 \\ \text{Cov}\left(\mu_0(X), \frac{(1-T)(Y - \mu_0(X))}{1-e(X)}\right) &= 0 \end{aligned}$$

Therefore we get that:

$$V_{\text{RR,AIPW}} = \tau_{\text{RR}}^2 \left(\text{Var}\left(\frac{\mu_1(X)}{\mathbb{E}[Y^{(1)}]} - \frac{\mu_0(X)}{\mathbb{E}[Y^{(0)}]}\right) + \mathbb{E}\left[\frac{\text{Var}(Y^{(1)}|X)}{e(X)\mathbb{E}[Y^{(1)}]^2}\right] + \mathbb{E}\left[\frac{\text{Var}(Y^{(0)}|X)}{(1-e(X))\mathbb{E}[Y^{(0)}]^2}\right] \right)$$

To conclude,

$$\sqrt{n}(\hat{\tau}_{\text{AIPW}} - \tau_{\text{RR}}) = \underbrace{\sqrt{n}(\tilde{\tau}_{\text{RR,AIPW}} - \tau_{\text{RR,AIPW}}^*)}_{\xrightarrow{p} 0} + \underbrace{\sqrt{n}(\tau_{\text{RR,AIPW}}^* - \tau_{\text{RR}})}_{\xrightarrow{d} \mathcal{N}(0, V_{\text{RR,AIPW}}x)}$$

where

$$V_{\text{RR,AIPW}} = \tau_{\text{RR}}^2 \left(\text{Var}\left(\frac{\mu_1(X)}{\mathbb{E}[Y^{(1)}]} - \frac{\mu_0(X)}{\mathbb{E}[Y^{(0)}]}\right) + \mathbb{E}\left[\frac{\text{Var}(Y^{(1)}|X)}{e(X)\mathbb{E}[Y^{(1)}]^2}\right] + \mathbb{E}\left[\frac{\text{Var}(Y^{(0)}|X)}{(1-e(X))\mathbb{E}[Y^{(0)}]^2}\right] \right)$$

□

A.2.3 Ratio one-step estimator

Proof of Definition 6. We will use Kennedy (2015) notation in this document. If you are not familiar on how to compute an influence function, note that it is very similar to compute the derivative of a function. We define our estimand quantity

$$\psi = \frac{\mathbb{E}[\mathbb{E}[Y|T=1, X]]}{\mathbb{E}[\mathbb{E}[Y|T=0, X]]} = \frac{\psi_1}{\psi_0}$$

We can now compute the influence function φ of ψ .

$$\begin{aligned} \varphi &= \mathbb{IF}(\psi) = \mathbb{IF}\left(\frac{\psi_1}{\psi_0}\right) = \frac{\mathbb{IF}(\psi_1)\psi_0 - \mathbb{IF}(\psi_0)\psi_1}{\psi_0^2} \\ &= \frac{\mathbb{IF}(\psi_1)}{\psi_0} - \psi \frac{\mathbb{IF}(\psi_0)}{\psi_0} \end{aligned}$$

We rely on Kennedy (2015) to show:

$$\begin{aligned} \mathbb{IF}(\psi_1) &= \mu_1(X) + T \frac{Y - \mu_1(X)}{e(X)} - \psi_1 \\ \mathbb{IF}(\psi_0) &= \mu_0(X) + (1-T) \frac{Y - \mu_0(X)}{1-e(X)} - \psi_0 \end{aligned}$$

Therefore, we have that:

$$\begin{aligned} \varphi &= \frac{\mathbb{IF}(\psi_1)}{\psi_0} - \psi \frac{\mathbb{IF}(\psi_0)}{\psi_0} \\ &= \frac{\mu_1(X) + T \frac{Y - \mu_1(X)}{e(X)} - \psi_1}{\psi_0} - \psi \frac{\mu_0(X) + (1-T) \frac{Y - \mu_0(X)}{1-e(X)} - \psi_0}{\psi_0} \\ &= \frac{\mu_1(X) + T \frac{Y - \mu_1(X)}{e(X)}}{\psi_0} - \psi - \psi \left(\frac{\mu_0(X) + (1-T) \frac{Y - \mu_0(X)}{1-e(X)}}{\psi_0} - 1 \right) \\ &= \frac{\mu_1(X) + T \frac{Y - \mu_1(X)}{e(X)}}{\psi_0} - \psi \frac{\mu_0(X) + (1-T) \frac{Y - \mu_0(X)}{1-e(X)}}{\psi_0} \end{aligned}$$

As referenced in Kennedy (2015) concerning the semiparametric von Mises expansion, consider the functional $\psi : P \rightarrow \mathbb{R}$, where P represents the true data distribution and \hat{P} its estimation. The expansion is formulated as:

$$\psi(\hat{P}) - \psi(P) = \int \varphi(z; \hat{P}) d(\hat{P} - P)(z) + R_2(\hat{P}, P),$$

applicable for distributions \hat{P} and P . The influence function $\varphi(z; P)$, associated with ψ , is a function with zero mean and finite variance as defined by Tsiatis:

$$\int \varphi(z; P) dP(z) = 0 \quad \text{and} \quad \int \varphi(z; P)^2 dP(z) < \infty,$$

and $R_2(\hat{P}, P)$ denotes a second-order remainder term, implying its reliance on higher degree differences between \hat{P} and P . This leads to the conclusion that standard plug-in estimators $\psi(\hat{P})$, are biased to the first order, evidenced by:

$$\psi(P) = \psi(\hat{P}) + \int \varphi(z; \hat{P}) dP(z) + R_2(\hat{P}, P),$$

while recognizing that:

$$\int \varphi(z; \hat{P}) d\hat{P}(z) = 0.$$

Therefore, we can build the one step estimator of $\psi(P)$ is:

$$\begin{aligned} \psi(P) &= \psi(\hat{P}) + \int \varphi(z; \hat{P}) dP(z) + R_2(\hat{P}, P) \\ &\approx \hat{\psi} + \frac{1}{n} \sum_{i=1}^n \varphi(Z_i) \\ &= \hat{\psi} + \frac{1}{n} \sum_{i=1}^n \frac{\mu_1(X_i) + T_i \frac{Y_i - \mu_1(X_i)}{e(X_i)}}{\hat{\psi}_0} - \hat{\psi} \frac{\mu_0(X_i) + (1 - T_i) \frac{Y_i - \mu_0(X_i)}{1 - e(X_i)}}{\hat{\psi}_0} \\ &= \hat{\psi} \left(1 - \frac{\frac{1}{n} \sum_{i=1}^n \mu_0(X_i) + (1 - T_i) \frac{Y_i - \mu_0(X_i)}{1 - e(X_i)}}{\hat{\psi}_0} \right) + \frac{\frac{1}{n} \sum_{i=1}^n \mu_1(X_i) + T_i \frac{Y_i - \mu_1(X_i)}{e(X_i)}}{\hat{\psi}_0} \\ \hat{\tau}_\varphi &:= \frac{\sum_{i=1}^n \hat{\mu}_1(X_i)}{\sum_{i=1}^n \hat{\mu}_0(X_i)} \left(1 - \frac{\sum_{i=1}^n \hat{\mu}_0(X_i) + \frac{(1 - T_i)(Y_i - \hat{\mu}_0(X_i))}{1 - \hat{e}(X_i)}}{\sum_{i=1}^n \hat{\mu}_0(X_i)} \right) + \frac{\sum_{i=1}^n \hat{\mu}_1(X_i) + \frac{T_i(Y_i - \hat{\mu}_1(X_i))}{\hat{e}(X_i)}}{\sum_{i=1}^n \hat{\mu}_0(X_i)} \end{aligned}$$

□

Proof of Proposition 6.

Asymptotic bias and variance of the crossfitted One-step estimator: Let us begin by examining the expression $\sqrt{n}(\hat{\tau}_\varphi - \tau_{RR})$. We express $\psi(P)$ as follows:

$$\psi(P) = \psi(\hat{P}) + \int \varphi(z; \hat{P}) dP(z) + R_2(\hat{P}, P),$$

where R_2 encapsulates higher order remainder terms.

To elucidate, we rearrange to find $\psi(\hat{P}) - \psi(P)$:

$$\begin{aligned} \psi(\hat{P}) - \psi(P) &= R_2(P, \hat{P}) - \int \varphi(z; \hat{P}) dP(z) \\ &= \frac{1}{n} \sum_{i=1}^n \varphi(z_i; P) - \frac{1}{n} \sum_{i=1}^n \varphi(z_i; \hat{P}) \\ &\quad + \frac{1}{n} \sum_{i=1}^n \left(\varphi(z_i; \hat{P}) - \varphi(z_i; P) \right) - \int \left(\varphi(z; \hat{P}) - \varphi(z; P) \right) dP(z) \\ &\quad + R_2(P, \hat{P}). \end{aligned}$$

Expanding on the difference $\tilde{\tau}_\varphi - \tau_{\text{RR}}$, we find:

$$\begin{aligned}\tilde{\tau}_\varphi - \tau_{\text{RR}} &= \psi(\hat{P}) + \frac{1}{n} \sum_{i=1}^n \varphi(z_i; \hat{P}) - \psi(P) \\ &= \frac{1}{n} \sum_{i=1}^n \varphi(z_i; P) \\ &\quad + \frac{1}{n} \sum_{i=1}^n \left(\varphi(z_i; \hat{P}) - \varphi(z_i; P) \right) - \int \left(\varphi(z; \hat{P}) - \varphi(z; P) \right) dP(z) \\ &\quad + R_2(P, \hat{P}).\end{aligned}$$

The initial term is a simple sample average of i.i.d. terms from a fixed function. According to the central limit theorem, it converges to a normally distributed random variable with variance $\text{Var}(\varphi(Z))/n$.

Now, consider the term involving differences:

$$\frac{1}{n} \sum_{i=1}^n \left(\varphi(z_i; \hat{P}) - \varphi(z_i; P) \right) - \int \left(\varphi(z; \hat{P}) - \varphi(z; P) \right) dP(z) = (P_n - P)\{\varphi(Z; \hat{P}) - \varphi(Z; P)\},$$

where P_n denotes the empirical measure.

This part of the proof is inspired from Vaart (1998). Considering the random variable $U = \left(\mathbb{P}_n^{(k)} - P \right) (\hat{\varphi}(Z) - \varphi(Z))$, and conditioning on observing a fixed dataset \mathcal{I}_{-k} , the mean and variance of U are computed as follows:

$$\begin{aligned}\mathbb{E}[U \mid \mathcal{I}_{-k}] &= 0, \\ \text{Var}[U \mid \mathcal{I}_{-k}] &\leq \frac{1}{n_k} \|\hat{\varphi}^{-k} - \varphi\|^2.\end{aligned}$$

Applying Chebyshev's inequality, we derive:

$$\mathbb{P} \left(\sqrt{\frac{1}{n_k}} |(\mathbb{P}_n^{(k)} - P)(\hat{\varphi} - \varphi)| \geq a \right) \leq \frac{1}{a^2}.$$

This demonstrates stochastic boundedness:

$$(P_n - P)(\hat{\varphi} - \varphi) = O_P \left(\frac{\|\hat{\varphi} - \varphi\|}{\sqrt{n}} \right) = o_P(1/\sqrt{n}),$$

asserting \mathcal{L}^2 consistency of $\hat{\varphi}$.

The final term $R_2(P, \hat{P})$ diminishes as:

$$R_2(P, \hat{P}) = o_p \left(\frac{1}{\sqrt{n}} \right).$$

Conclusively, we establish that:

$$\sqrt{n}(\tilde{\tau}_\varphi - \tau_{\text{RR}}) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \varphi(z_i; P) + o_p \left(\frac{1}{\sqrt{n}} \right).$$

Thus, we find:

$$\sqrt{n}(\hat{\tau}_\varphi - \tau_{\text{RR}}) \xrightarrow{d} \mathcal{N}(0, \text{Var}(\varphi)),$$

where $\text{Var}(\varphi) = V_{\text{RR,AIPW}}$. □

B Simulation

For the simulations we have implemented all estimators in Python using Scikit-Learn for our regression and classification models. All our experiments were run on a 8GB M1 Mac. The propensity scores is estimated based on the provided training data and covariate names. Depending on the chosen method, it either uses logistic regression with a high regularization parameter (parametric) or a random forest classifier with parameters determined by the training data size (non-parametric). The response surface is estimated based on the training data, covariate names, the method (parametric or non-parametric), and whether the response is binary or continuous. For parametric methods, it uses a stochastic gradient descent classifier for binary responses and a linear regression model for continuous responses. For non-parametric methods, it employs a random forest classifier for binary responses and a random forest regressor for continuous responses. Both methods fit the model using the training data to estimate the respective scores and surfaces, enabling flexible handling of various datasets and assumptions for causal inference analysis.

B.1 Randomized Controlled Trials

In this part we will simulate Randomized Controlled Trials (RCT) and test the following Ratio estimators: Ratio Neyman, Ratio Horvitz Thomson and the Ratio G-formula. Since we are in a Randomized Controlled Trials, the propensity score $e(\cdot)$ is constant.

B.1.1 Linear RCT

The first DGP has linear outcome models (linear treatment effect and the baseline). The data is generated using:

$$\begin{aligned} m(X) &= (c_1 - c_0) + (\beta_1 - \beta_0)^\top X & c_0 &= 6, & c_1 &= 12 \\ b(X) &= c_0 + \beta_0^\top X & \beta_1 &= (2, -5, 2, 8, -2, 8) \\ e(X) &= 0.5 & \beta_0 &= (3, -7, 1, 4, -2, 2) \end{aligned}$$

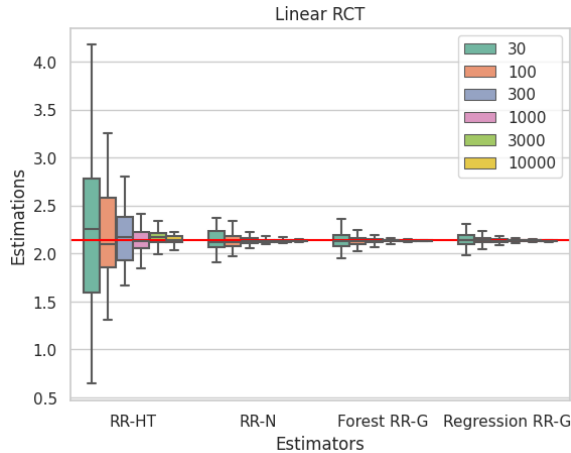


Figure 4: Comparison of RCT estimators in a Linear RCT

Given that X has a zero mean, it follows that $\tau_{RR} = c_1/c_0 = 2$. This scenario aligns with the linear setting outlined in 28. Referring to Figure 4, as proved in the previous sections all estimators converge to the true Risk Ratio as n increases. Additionally, within this linear framework as per Lemma 1, the variance of the Neyman estimator exceeds the one of the G-formula. In such a linear environment, the parametric G-formula performs better than its non-parametric counterpart. Additionally, the Ratio Neyman estimator demonstrates lower variance compared to the Horvitz-Thomson estimator as indicated in Equation (5).

B.1.2 Non-Linear RCT

This DGP is also a Randomized Controlled Trials however, the outcomes are not linear this time:

$$m(X) = \sin(X_1) \cdot X_2^2 + \frac{X_3}{X_4 + 1} - \log(X_5 + 1) + X_6^3 + 1$$

$$b(X) = 4 * \max(X_1 + X_2 + X_3, 0) - \min(X_4 + X_6, 0) \quad \text{and} \quad e(X) = 0.5$$

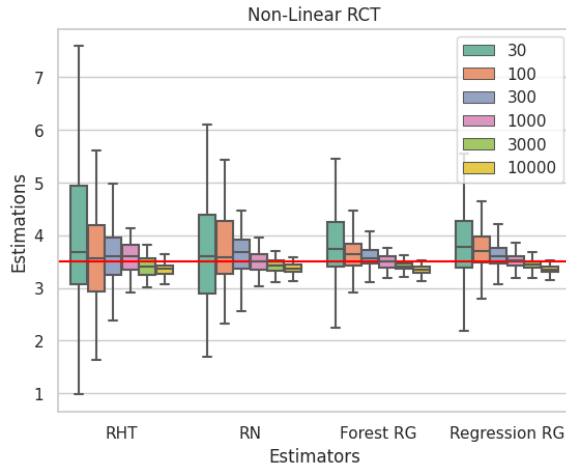


Figure 5: Comparison of RCT estimators in a Non-Linear RCT

The presence of trigonometric, exponential, logarithmic, and polynomial terms makes this setting non-linear. It's important to note that since we are in a Randomized Controlled Trial (RCT), the propensity function remains constant. As the sample size (n) increases, all proposed estimators converge. A bias can be seen in 5 but decreases to 0 as (n) increases as predicted in previous sections. Linear regression struggles with small n values, failing to capture the intricate relationships between features and non-linearities. On the other hand, Random Forest, a non-parametric method, excels in capturing these complexities by segmenting the feature space and predicting based on response averages within those segments. However, predicting the complex function can be challenging, the Neyman estimator might outperform the G-formula, particularly when both parametric and non-parametric responses may lack consistency. Although we do not fall in assumptions of Equation (5) the Ratio Neyman estimator demonstrates lower variance compared to the Horvitz-Thomson estimator.

NeurIPS Paper Checklist

1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: [Yes]

Justification: All the claims presented in the abstract and introduction are formalized as propositions within the paper, based on classical causal inference assumptions.

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [Yes]

Justification: We discuss (mainly in the conclusion) the limitation of our work.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. Theory Assumptions and Proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: Rigorous proofs are provided in the Appendix, with references to the theorems and lemmas upon which our proofs rely.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. Experimental Result Reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We described in the paper or appendix the settings of the model we use and parameters we used for our code.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
 - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
 - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
 - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
 - (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: Experiments are considered simple enough to be reproduced without providing open access to our code.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. Experimental Setting/Details

Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: Experimental Setting/Details are fully provided in supplemental material

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. Experiment Statistical Significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: We used boxplots to report our estimations.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. Experiments Compute Resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: Yes, the paper does indicate the type of compute used for the experiments.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. Code Of Ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines?>

Answer: [Yes]

Justification: We made sure to preserve anonymity in our paper.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. Broader Impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: We discussed the impact of our findings to better estimate treatment effects using non-randomized data.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: The paper poses no such risks

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: All theorems used are properly credited.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset’s creators.

13. **New Assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: The paper does not release new assets

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.