University of California, Berkeley

U.C. Berkeley Division of Biostatistics Working Paper Series

Year 2008 *Paper* 229

Covariate Adjustment for the Intention-to-Treat Parameter with Empirical Efficiency Maximization

Daniel B. Rubin* Mark J. van der Laan†

*Division of Biostatistics, School of Public Health, University of California, Berkeley, daniel.rubin@fda.hhs.gov

†Division of Biostatistics and Department of Statistics, University of California, Berkeley, laan@berkeley.edu

Covariate Adjustment for the Intention-to-Treat Parameter with Empirical Efficiency Maximization

Daniel B. Rubin and Mark J. van der Laan

Abstract

In randomized experiments, the intention-to-treat parameter is defined as the difference in expected outcomes between groups assigned to treatment and control arms. There is a large literature focusing on how (possibly misspecified) working models can sometimes exploit baseline covariate measurements to gain precision, although covariate adjustment is not strictly necessary. In Rubin and van der Laan (2008), we proposed the technique of empirical efficiency maximization for improving estimation by forming nonstandard fits of such working models. Considering a more realistic randomization scheme than in our original article, we suggest a new class of working models for utilizing covariate information, show our method can be implemented by adding weights to standard regression algorithms, and demonstrate benefits over existing estimators through numerical asymptotic efficiency calculations and simulations.

1 The Intention-to-Treat Parameter

Consider an experiment in which baseline measurements are taken on n subjects, a random subsample of size m_n is assigned to a treatment, the remaining $n-m_n$ subjects are assigned to a control arm, and outcomes are later assessed. The observed data is

$$\{O_i\}_{i=1}^n = \{W_i, \Delta_i, Y_i\}_{i=1}^n,$$

where W_i is the i^{th} subject's covariate vector, treatment or control is indicated by $\{\Delta_i = 1\}$ or $\{\Delta_i = 0\}$, and $Y_i \in \mathbb{R}$ is the outcome. Suppose interest lies in estimating the treatment effect

$$\mu = E[Y|\Delta = 1] - E[Y|\Delta = 0] = \mu_T - \mu_C. \tag{1}$$

This difference in means between those assigned to treatment and control is often called the intention-to-treat parameter. For a binary response, this value can be termed the excess risk or risk difference.

In the counterfactual outcome formulation of Neyman (1923) or Rubin (1974), the unavailable full data would be

$$\{X_i\}_{i=1}^n = \{W_i, Y_{i,T}, Y_{i,C}\}_{i=1}^n,$$

where $Y_{i,T}$ and $Y_{i,C}$ denote the responses that would have occurred for the i^{th} subject under treatment and control. Because only one of these counterfactual outcomes is ever seen, the observed response is

$$Y_i = \Delta_i Y_{i,T} + (1-\Delta_i) Y_{i,C},$$

and the treatment effect (1) can then be written as $\mu = E[Y_T] - E[Y_C] = \mu_T - \mu_C$. Throughout this work, we assume

$$\text{the full data } \{X_i\}_{i=1}^n = \{W_i, Y_{i,T}, Y_{i,C}\}_{i=1}^n \text{ would consititute an i.i.d. sample.} \tag{2}$$

Note that the observed $O_1, ..., O_n$ cannot be independent, because the Δ_i's must sum to the number m_n of subjects assigned treatment. However, $O_1, ..., O_n$ are identically distributed, and we write $O = (W, \Delta, Y)$ as a random variable drawn from this common distribution. In this nonparametric setting, we will additionally assume that

$$\text{response } Y \text{ has finite variance, implied by } E[Y_T^2] < \infty \text{ and } E[Y_C^2] < \infty. \tag{3}$$

A simple approach would be to ignore covariates and estimate μ with the difference in sample means

$$\mu_n = \frac{1}{m_n} \sum_{\{i:\ \Delta_i=1\}} Y_i - \frac{1}{n-m_n} \sum_{\{i:\ \Delta_i=0\}} Y_i. \tag{4}$$

Such an estimator would be unbiased, and in large samples $\sqrt{n}(\mu_n - \mu)$ would be approximately Gaussian with mean zero, so this μ_n would in fact be perfectly valid. However, it has been recognized at least since Fisher (1932) that ignoring informative covariates is potentially wasteful, because covariate W_i might inform how the i^{th} subject would have responded in both the treatment and control arms.

2 Covariate Adjustment

Proper adjustment is straightforward when covariates naturally partition the subjects into a handful of strata, but requires thought with even a single continuous explanatory variable such as age, let alone when modern studies collect copious amounts of baseline information. Pocock et al. (2002) surveyed 50 clinical trial reports, and found that 36 used covariate adjustment, and that 12 reports emphasized adjusted over unadjusted analysis. The authors remarked that "Nevertheless, the statistical emphasis on covariate adjustment is quite complex and often poorly understood, and there remains confusion as to what is an appropriate statistical strategy."

Letting $\pi_n = \frac{1}{n}\sum_{i=1}^{n}\Delta_i = \frac{m_n}{n}$ denote the proportion of subjects assigned to treatment, we will examine estimators of the form

$$\mu_{n,Q} = \frac{1}{n}\sum_{i=1}^{n}(\frac{\Delta_i}{\pi_n} - \frac{1-\Delta_i}{1-\pi_n})(Y_i - Q(W_i)), \tag{5}$$

and estimators corresponding to using wise fits $Q_n(\cdot) = Q_n(\cdot;\ O_1, ..., O_n)$ built from the data. When $Q(\cdot)$ is constant, the estimator (5) reduces to the unadjusted difference in means (4). We will consider functions Q: Support$(W) \to \mathbb{R}$ such that

$$Q(W) \text{ is measurable and has finite variance.} \tag{6}$$

The functions $Q_n(\cdot;\ O_1, ..., O_n)$ are assumed to fall in such a class with probability one.

The following lemma summarizes the behavior of estimator $\mu_{n,Q}$, and all proofs are deferred to the appendix.

Lemma 1. *For $0 < m_n < n$, the estimator $\mu_{n,Q}$ of $\mu = \mu_T - \mu_C$ is unbiased. Further, $\sqrt{n}(\mu_{n,Q} - \mu)$ has variance*

$$\begin{aligned} \sigma^2(Q,\pi_n) &\equiv \frac{1}{\pi_n}E|Y_T - Q(W)|^2 + \frac{1}{1-\pi_n}E|Y_C - Q(W)|^2 \\ &\quad -\frac{1}{\pi_n}(\mu_T - E[Q(W)])^2 - \frac{1}{1-\pi_n}(\mu_C - E[Q(W)])^2 \\ &= E[(\frac{\Delta}{\pi_n^2} + \frac{1-\Delta}{(1-\pi_n)^2})|Y - Q(W)|^2] \\ &\quad -\frac{1}{\pi_n}|\mu_T - E[Q(W)]|^2 - \frac{1}{1-\pi_n}|\mu_C - E[Q(W)]|^2. \end{aligned}$$

Consider the sequence of experiments in which covariates $\{W_i\}_{i=1}^n$ are measured, a random subsample of m_n subjects are assigned to treatment and the remaining $n - m_n$ to the control arm. If $\pi_n = \frac{m_n}{n} \to \pi$ and $0 < \pi < 1$, then $\sqrt{n}(\mu_{n,Q} - \mu)$ converges in law to a Gaussian distribution with mean zero and variance $\sigma^2(Q,\pi)$.

Note that the value $\sigma^2(Q,\pi)$ is determined by the function $Q(\cdot)$, scalar π, and distribution of $\{W, Y_T, Y_C\}$, so is well-defined even if $\pi_n \neq \pi$ for a certain sample size.

When μ_{n,Q_n} is applied with a function $Q_n(\cdot;\ O_1, ..., O_n)$ built from the data, the estimator might no longer be exactly unbiased. However, the next lemma tells us that as long as Q_n converges to some function Q_0 in a certain sense, the $\sqrt{n}$-asymptotics will be as if the limiting function were known and we applied μ_{n,Q_0}. It will not even be necessary for the $Q_n \to Q_0$ convergence to occur at any rate.

Lemma 2. *Consider the sequence of experiments in which baseline covariates $\{W_i\}_{i=1}^n$ are measured, a random subsample of m_n subjects are assigned to treatment and the remaining $n - m_n$ to the control arm. Let $\pi_n = \frac{m_n}{n} \to \pi$, for which $0 < \pi < 1$. Suppose that $Q_n(\cdot) = Q_n(\cdot;\ O_1, ..., O_n)$ is a random function determined by the observed data, mapping Support(W) to $\mathbb{R}$. For simplicity, we will consider $Q_n(\cdot)$ not depending on how the observations are ordered, so $Q_n(\cdot;\ O_1, ..., O_n) = Q_n(\cdot;\ O_{i_1}, ...O_{i_n})$ for any permutation $(i_1, ..., i_n)$ of $(1, ..., n)$. Letting P_W denote the marginal distribution of covariate W, assume there is a $Q_0(\cdot)$ such that*

$$\int |Q_n(w) - Q_0(w)|^2 dP_W(w) \to 0 \text{ in probability.} \tag{7}$$

Assume further that

$$\begin{gathered}\text{there is a } P_W\text{-Donsker class of functions } \mathcal{Q}_0 \text{ such that} \\ Q_n \text{ falls in } \mathcal{Q}_0 \text{ with probability tending to one.}\end{gathered} \tag{8}$$

Define $\mu_{n,Q_n} = \frac{1}{n}\sum_{i=1}^n (\frac{\Delta_i}{\pi_n} - \frac{1-\Delta_i}{1-\pi_n})(Y_i - Q_n(W_i))$. Then

$$\sqrt{n}(\mu_{n,Q_n} - \mu) \to N(0, \sigma^2(Q_0, \pi)) \text{ in law.}$$

We will soon see that both recent and classical procedures correspond to using estimators μ_{n,Q_n}, and attempting to exploit covariate information through working models for the data generating distribution.

3 What is a Good Q?

The preceding lemma reveals that the asymptotic performance of our parameter estimate will be determined by the function Q_0 to which Q_n converges. The following lemma may help clarify when this limit Q_0 will lead to a small asymptotic variance.

Lemma 3. *Let*

$$\begin{aligned} MSE(Q, \pi_n) &= \sigma^2(Q, \pi_n) + \frac{1}{\pi_n}|\mu_T - E[Q(W)]|^2 + \frac{1}{1-\pi_n}|\mu_C - E[Q(W)]|^2 \\ &= E[(\frac{\Delta}{\pi_n} - \frac{1-\Delta}{1-\pi_n})^2 |Y - Q(W_i)|^2] \end{aligned}$$

denote a weighted mean squared error of $Q(W)$ as a prediction of response Y. The following relationships hold, where "minimizes" does not necessarily mean "uniquely minimizes."

(A) For constant c, we have that $\sigma^2(Q + c, \pi_n) = \sigma^2(Q, \pi_n)$, meaning that adding a constant to function Q won't affect variance of $\mu_{n,Q}$ (because it won't affect the estimator for any sample size).

(B) *Over constants c, the value $c^\star = \pi_n\mu_T + (1-\pi_n)\mu_C - E[Q]$ minimizes the weighted mean squared error $MSE(Q+c, \pi_n)$.*

(C) *If $MSE(Q, \pi_n) \le MSE(\pi_n\mu_T + (1-\pi_n)\mu_C, \pi_n)$, then $\sigma^2(Q, \pi_n) \le \sigma^2(c, \pi_n)$. Thus, the variance of estimator $\mu_{n,Q}$ is no larger than the variance of the unadjusted estimator (4) making no use of covariate information. By the previous statement (B) applied with $Q(W) = 0$, this condition holds if and only if $MSE(Q, \pi_n) \le MSE(c, \pi_n)$ for all constants c.*

(D) *Let $\mathcal{Q}_1 = \{Q + c : Q \in \mathcal{Q}, c \in \mathbb{R}\}$ be the expanded function class corresponding to all shifts of functions in class $\mathcal{Q}$. If $Q_1 = Q_0 + c$ minimizes the weighted mean squared error $MSE(Q, \pi_n)$ over $\mathcal{Q}_1$, then Q_0 minimizes variance $\sigma^2(Q, \pi_n)$ over $\mathcal{Q}$. In particular, if $\mathcal{Q}$ is closed under shifts then minimizing $MSE(Q, \pi_n)$ over $\mathcal{Q}$ corresponds to minimizing $\sigma^2(Q, \pi_n)$ over $\mathcal{Q}$.*

(E) *The function $Q^\star$ minimizing variance $\sigma^2(Q, \pi_n)$ over all functions Q is given by the regression of $Y^\star = [\frac{(1-\pi_n)\Delta}{\pi_n} + \frac{\pi_n(1-\Delta)}{1-\pi_n}]Y$ on W, which is*

$$\begin{aligned} Q^\star(W) &= (1-\pi_n)E[Y|\Delta = 1, W] + \pi_n E[Y|\Delta = 0, W] \\ &= (1-\pi_n)E[Y_T|W] + \pi_n E[Y_C|W]. \end{aligned}$$

The results demonstrate that the asymptotic variance $\sigma^2(Q, \pi)$ is closely related to the weighted mean squared error $MSE(Q, \pi)$. If a function class $\mathcal{Q}$ includes all constant functions, minimizing the mean squared error over the function class can only lead to improvement relative to the unadjusted estimator. Moreover, if the function class $\mathcal{Q}$ is closed under shifts, minimizing the weighted mean squared error corresponds to minimizing the desired asymptotic variance.

4 Empirical Efficiency Maximization

Lemma 3 identifies the optimal $Q^\star$ for use in the parameter estimate, which unfortunately depends on unknown nuisance parameters. Although we wouldn't know the regression function of $Y^\star = [\frac{(1-\pi_n)\Delta}{\pi_n} + \frac{\pi_n(1-\Delta)}{1-\pi_n}]Y$ on covariate W, we could posit a class of functions $\mathcal{Q}$, and hope that at least one element $Q \in \mathcal{Q}$ might well approximate $Q^\star$.

The function class $\mathcal{Q}$ should be chosen to include all constant functions, so that the element minimizing $MSE(Q, \pi_n)$ will not be asymptotically inferior to the unadjusted estimator. The class should be expanded to be closed under shifts, so the element minimizing the weighted mean squared error also minimizes asymptotic variance.

The weighted mean squared error $MSE(Q, \pi_n)$ can then be used to determine an appropriate $Q \in \mathcal{Q}$, as it can be approximated empirically with the unbiased

$$MSE_n(Q, \pi_n) = \frac{1}{n}\sum_{i=1}^{n}(\frac{\Delta_i}{\pi_n} - \frac{1-\Delta_i}{1-\pi_n})^2|Y_i - Q(W_i)|^2.$$

For a relatively large sample size n and relatively small function class $\mathcal{Q}$, we might hope the empirical $MSE_n(Q, \pi_n)$ approximates the true weighted mean squared error

$MSE(Q, \pi_n)$ uniformly over $\mathcal{Q}$. The empirical minimizer might then approximate the population minimizer. Our proposal is thus to select

$$Q_n = \text{argmin}_{Q \in \mathcal{Q}} MSE_n(Q, \pi_n), \tag{9}$$

and estimate the parameter of interest with $\mu_{n,Q_n} = \frac{1}{n}\sum_{i=1}^{n}(\frac{\Delta_i}{\pi_n} - \frac{1-\Delta_i}{1-\pi_n})^2(Y_i - Q_n(W_i))$.

Of course, the empirical minimizer might not exist for a poorly chosen function class $\mathcal{Q}$, and even if it does exist it might not be unique. For nonlinear function classes $\mathcal{Q}$, numerical considerations might force one to settle for a local minimizer.

If we were regressing response Y on covariates W, solving (9) would correspond to fitting the model with least squares, and observation weights $(\frac{\Delta_i}{\pi_n} - \frac{1-\Delta_i}{1-\pi_n})^2$. That is, each observation in the treatment group would be given a weight proportional to π_n^{-2}, and each observation in the control group would be given a weight proportional to $(1-\pi_n)^{-2}$. Therefore, if an algorithm exists to solve the least squares problem over the function class $\mathcal{Q}$, and accepts observation weights, empirical efficiency maximization can be easily implemented.

Whatever optimization algorithm is used to form Q_n as in (9), the previously given Lemma 2 tells us asymptotic Gaussianity can be achieved if $\int |Q_n(w) - Q_0(w)|^2 dP_W(w)$ converges to zero in probability, for some function Q_0. We might justifiably anticipate such a Q_0 to have a small weighted mean squared error $MSE(Q, \pi_n)$, which ideally would be close to $\inf_{Q \in \mathcal{Q}} MSE(Q, \pi_n)$, and consequently lead to a precise estimator for the intention-to-treat parameter.

We discuss empirical efficiency maximization more extensively in Rubin and van der Laan (2008). There we also consider estimation of (log) relative risks and (log) odds ratios in randomized experiments, as well as the expected outcomes μ_T and μ_C in the treatment and control groups themselves. The method also has applicability in survival analysis, and in general coarsened data structures when the coarsening mechanism can be correctly modeled.

4.1 Continuous Outcome

With a continuous outcome Y, one might initially think to approximate the outcome regression of $Y^\star = [\frac{(1-\pi_n)\Delta}{\pi_n} + \frac{\pi_n(1-\Delta)}{1-\pi_n}]Y$ with a linear model, inducing the function class

$$\mathcal{Q} = \{Q_{\alpha,\beta}(w) = \alpha + \beta^T w : (\alpha, \beta)\}.$$

We would then solve

$$(\alpha_n, \beta_n) = \text{argmin}_{\alpha,\beta} \frac{1}{n}\sum_{i=1}^{n}(\frac{\Delta_i}{\pi_n} - \frac{1-\Delta_i}{1-\pi_n})^2(Y_i - \alpha - \beta^T W_i),$$

with weighted linear least squares, and apply μ_{n,Q_n} with $Q_n(w) = \alpha_n + \beta_n^T w$. Section 7 contrasts this approach with the more traditional ANCOVA method of performing covariate adjustment through linear modeling.

Tsiatis et al. (2000) further discusses estimators corresponding to $\mu_{n,Q}$ with linear functions Q. An estimator due to Koch et al. (1998) is shown to be asymptotically optimal among such estimators, so in this special case empirical efficiency maximization actually adds no novel methodology.

The story changes when one attempts to perform more aggressive covariate adjustment as in Tsiatis in et al., as we will argue in the following section contrasting empirical efficiency maximization with standard locally efficient estimation. Considering the prediction of response Y from covariates W, one might employ techniques such as covariate transformations, variable/model selection, cross-validation, penalization, nonlinear least squares, partitioning, additive models, local linear fits, partitioning, or a number of other methods. Present results reveal these prediction algorithms should be applied with observation weights proportional to π_n^{-2} and $(1-\pi_n)^{-2}$ in the treatment and control groups. As long as some $Q_n(\cdot) \to Q_0(\cdot)$ convergence can be ensured, Lemma 2 tells us model misspecification will not compromise consistency and asymptotic Gaussianity. All stages of model building should be targeted toward selecting a function $Q(\cdot)$ leading to a desirable intention-to-treat estimate, and this can be enhanced through simple observation weighting.

4.2 Binary Outcome

With a binary outcome $Y \in \{0,1\}$, such as an outcome corresponding to the presence or absence of disease, empirical efficiency maximization diverges more dramatically from existing methods for covariate adjustment. The intention-to-treat parameter is in this case the difference of disease probabilities in the treatment and control groups.

Recall that the optimal function Q is the regression of $Y^\star = [\frac{(1-\pi_n)\Delta}{\pi_n} + \frac{\pi_n(1-\Delta)}{1-\pi_n}]Y$ on W, given by $Q^\star(W) = (1-\pi_n)E[Y|\Delta = 1, W] + \pi_n E[Y|\Delta = 0, W]$. One initial thought might be to fit a logistic regression for $P(Y = 1|\Delta, W)$, and substitute into $Q^\star(\cdot)$ accordingly. Another thought, although we haven't seen this proposed, would be a multinomial logit model for the regression of $Y^\star \in \{0, \frac{1-\pi_n}{\pi_n}, \frac{\pi_n}{1-\pi_n}\}$ on covariates W.

Such approaches will lead to asymptotic efficiency when using correctly specified working models. However, with incorrect logistic or multinomial logit models, estimation can suffer. The reason is that a working model induces a function class $\mathcal{Q}$. Likelihood-based methods aim to minimize Kullback-Leibler divergence from the truth, not select the optimal $Q \in \mathcal{Q}$ for the parameter of interest. We showed in Rubin and van der Laan (2008) how a likelihood-based logistic regression approach can be inferior to unadjusted estimation.

Because $0 \leq Q^\star(W) = E[Y^\star|W] \leq 1$, we might consider approximating this optimal function with a sigmoid in a linear combination of the covariates. Expanding to ensure closure under shifts, the working function class becomes

$$\mathcal{Q} = \{Q_{c,\alpha,\beta}(w) = c + \frac{1}{1+\exp(-\alpha - \beta^T w)} : (c, \alpha, \beta)\}.$$

The empirical efficiency maximization approach would then be to perform a weighted nonlinear least squares to choose

$$Q_n = \text{argmin}_{Q_{c,\alpha,\beta}} MSE_n(Q_{c,\alpha,\beta}, \pi_n) = \text{argmin}_{Q_{c,\alpha,\beta}} \frac{1}{n}\sum_{i=1}^{n}(\frac{\Delta_i}{\pi_n} - \frac{1-\Delta_i}{1-\pi_n})^2 |Y_i - Q_{c,\alpha,\beta}(W_i)|^2,$$

and estimate the parameter of interest with $\mu_{n,Q_n} = \frac{1}{n}\sum_{i=1}^{n}(\frac{\Delta_i}{\pi_n} - \frac{1-\Delta_i}{1-\pi_n})(Y_i - Q_n(W_i))$.

Note that the parameters (c, α, β) aren't identified by the function $Q_{c,\alpha,\beta}$, as constant functions can be parametrized in several ways when $\beta = 0$. So long as we have a $Q_n \rightarrow Q_0$ convergence, it is irrelevant in Lemma 2 whether we have a (c_n, α_n, β_n) convergence, because the intention-to-treat estimate μ_{n,Q_n} does not depend on how Q_n is parametrized.

A potential difficulty is that the optimization problem can no longer be solved in closed form. We use the **nlminb()** function in R for computations in Sections 9 and 10. Note that the function class $\mathcal{Q}$ corresponds to a neural network with a single output neuron, in which the neuron is not scaled. Sontag and Sussmann (2001) assert

> It seems to be 'folk knowledge' that no spurious local minima can happen when there are no hidden neurons. (The argument made in the last case is roughly that the problem should be analogous to a standard least squares problem, in which neurons have a linear response map.)

Nevertheless, Sontag and Sussmann show the folk knowledge can be false, and that local minima can indeed occur.

We might expect our Q_n converges to a Q_0 corresponding to going "downhill" with gradient descent on $(c, \alpha, \beta) \rightarrow MSE(Q_{c,\alpha,\beta}, \pi_n)$, when starting from an initial value (c_0, α_0, β_0). If the stochastic process $(c, \alpha, \beta) \rightarrow MSE_n(Q_{c,\alpha,\beta}, \pi_n)$ converges to the function $(c, \alpha, \beta) \rightarrow MSE(Q_{c,\alpha,\beta}, \pi)$ in the Glivenko-Cantelli sense, the continuous mapping theorem could imply such a result, provided this "downhill" argmin functional is continuous with respect to the appropriate metric. See, for example, the discussion of M-estimators in Chapter 3.2 of van der Vaart and Wellner (1996). However, there may be pathological cases where solving for Q_n with the wrong optimization algorithm doesn't necessarily guarantee a $Q_n \rightarrow Q_0$ convergence as in Lemma 2, and hence doesn't guarantee asymptotic Gaussianity of the intention-to-treat estimator.

Work is ongoing to determine what modifications can be made, if any, to ensure a $Q_n \rightarrow Q_0$ convergence when fitting a logistic regression model with empirical efficiency maximization. Until then, our heuristic is that standard optimization algorithms should usually force some $Q_n \rightarrow Q_0$ convergence, for which the weighted mean squared error $MSE(Q_0, \pi)$ approximates $\inf_{Q \in \mathcal{Q}} MSE(Q, \pi)$. Lemmas 2 and 3 would then imply desirable $\sqrt{n}$-asymptotics, which itself is a heuristic for accuracy in finite samples.

5 Relationship to Locally Efficient Estimation

A general methodology for estimation in coarsened data structures was formulated in Robins and Rotnitzky (1992) and Robins, Rotnitzky, and Zhao (1994), and was primarily motivated by causal inference problems in observational studies. The approach is surveyed in the books of van der Laan and Robins (2003) and Tsiatis (2006).

Temporarily consider what happens when treatment is not randomized, but the observed data is an i.i.d. sample

$$\{O_i\}_{i=1}^{n} = \{W_i, \Delta_i, Y_i = \Delta_i Y_{T,i} + (1-\Delta_i) Y_{C,i}\},$$

where again W_i is a covariate vector, Δ_i is the observed treatment, and Y_i is the observed outcome. With the strong unverifiable assumption $\{(Y_T, Y_C) \perp \Delta | W\}$ of no unmeasured confounding, and the assumption that $P(\Delta = 1|W)$ is bounded away from zero and one, the intention-to-treat parameter is identifiable.

Let $g(\cdot)$, $Q_T(\cdot)$, and $Q_C(\cdot)$ be arbitrary functions, and consider the estimator

$$\mu_n = \frac{1}{n}\sum_{i=1}^{n}\{\frac{\Delta_i Y_i}{g(W_i)} + (1 - \frac{\Delta_i}{g(W_i)})Q_T(W_i) - \frac{(1-\Delta_i)Y_i}{1-g(W_i)} - (1 - \frac{1-\Delta_i}{1-g(W_i)})Q_C(W_i)\}. \quad (10)$$

It can be shown the estimator is unbiased if either $g(W) = P(\Delta = 1|W)$ or if $(Q_T(W), Q_C(W)) = (E[Y|\Delta = 1, W], E[Y|\Delta = 0, W])$. One can then fit working models for the treatment mechanism $\mathcal{L}(\Delta|W)$ and the outcome distribution $\mathcal{L}(Y|\Delta, W)$. For instance, the former might correspond with a logistic regression of treatment on covariates, and the latter with a linear regression of the outcome on treatment and covariates. Resulting model fits lead to $g(\cdot)$ and $(Q_T(\cdot), Q_C(\cdot))$ for substitution into the intention-to-treat estimate.

The procedure is usually advertised as being doubly robust and locally efficient. Double robustness means that only one of the treatment mechanism or outcome distribution models has to be correctly specified. Local efficiency means that if both working models are correctly specified, the parameter estimate will be asymptotically efficient, and achieve an asymptotic information bound.

In Rubin and van der Laan (2008), we pondered what these two beneficial properties provide for estimator (10) in randomized experiments. Double robustness is superfluous, as the treatment mechanism $P(\Delta_i|W_i) = \pi_n$ is known, and hence doesn't have to modeled. However, it can paradoxically be shown that fitting this known constant function with a correctly specified working model would only enhance efficiency.

Local efficiency tells us that if we can correctly model the regression function $E[Y|\Delta, W]$ then we will be asymptotically efficient. This can be done with simple averaging when covariate W is discrete and takes on a small number of levels such as when $W \in \{\text{Male, Female}\}$. But when positing a working model for the regression of outcome Y on treatment Δ and covariate W, local efficiency only tells us what happens when the model is correct. When the model is incorrect, we showed in Rubin and van der Laan (2008) the resulting estimator (10) can have larger asymptotic variance than the unadjusted estimator making no use of covariate information.

Our empirical efficiency maximization proposal was to fit the working model for the regression of Y on (Δ, W) not with maximum likelihood, but to ensure the limiting fits of $(E[Y|\Delta = 1, W], E[Y|\Delta = 0, W])$ minimized the asymptotic variance of the intention-to-treat estimator.

The simple insight of the present work is that $\mu_n = \mu_{n,(g,Q_T,Q_C)}$ of (10) is simply the estimator $\mu_{n,Q}$ of (5) with $g(W) = \pi_n$ and $Q(W) = (1 - \pi_n)Q_T(W) + \pi_n Q_C(W)$. Hence, a working model $\mathcal{F}$ for the outcome distribution $\mathcal{L}(Y|\Delta, W)$ induces a class of functions

$$\mathcal{Q} = \{w \rightarrow (1 - \pi_n)E_F[Y|\Delta = 1, W = w] + \pi_n E_F[Y|\Delta = 0, W = w] : F \in \mathcal{F}\}.$$

The optimal working model element $F \in \mathcal{F}$ for the asymptotic variance of the intention-to-treat estimator can then be found as in Section 4, through minimizing a weighted mean squared error.

When the working model is misspecified, the resulting estimator can be more efficient than the typical technique for locally efficient estimation, which is to fit the working model F with maximum likelihood. Likelihood-based estimation of $F \in \mathcal{F}$ might converge to the element minimizing a Kullback-Leibler divergence from the true data generating distribution of $\mathcal{L}(Y|\Delta, W)$, while empirical efficiency maximization attempts to find the working model element minimizing asymptotic variance for the resulting parameter estimate. Benefits over standard locally efficient methods are shown through asymptotic efficiency calculations and simulations in Section 9 and Section 10.

6 Alternative Modeling

Lemma 3 may provide guidance as to how to form a working model for covariate adjustment in the first place. Consider the case of a binary outcome $Y \in \{0, 1\}$. A standard locally efficient procedure would be to fit a logistic regression model

$$\text{logit } P(Y = 1|\Delta, W) = \alpha + \gamma\Delta + \beta^T W,$$

inducing the class of functions

$$\mathcal{Q} = \{Q_{\alpha,\gamma,\beta}(w) = \frac{1 - \pi_n}{1 + \exp(-\alpha - \gamma - \beta^T W)} + \frac{\pi_n}{1 + \exp(-\alpha - \beta^T W)} : (\alpha, \gamma, \beta)\}.$$

Weighted nonlinear least squares could be used to approximate the function class element leading to smallest asymptotic variance. However, recall that the optimal function is given by $Q^\star(W) = (1-\pi_n)E[Y|\Delta = 1, W] + \pi_n E[Y|\Delta = 0, W]$, which is between zero and one. It might make no less clinical sense for the regression function $Q^\star = E[Y^\star|W]$ to grow sigmoidally in a linear function of the covariates, as in Section 4.2.

It may often be more convenient to model the regression of $Y^\star = [\frac{(1-\pi_n)\Delta}{\pi_n} + \frac{\pi_n(1-\Delta)}{1-\pi_n}]Y$ on covariates W, rather than the regression of outcome Y on both the covariates and treatment (W, Δ). Equivalently, we can think of building a predictor of response Y from covariates W, but using observation weights $(\frac{\Delta_i}{\pi_n} - \frac{1-\Delta_i}{1-\pi_n})^2$.

One could also fit separate regression models for $E[Y|\Delta = 1, W]$ and $E[Y|\Delta = 0, W]$ in the treatment and control arms, but we see several disadvantages. Primarily, the optimal element in one working model can depend on the element being used in the other. Moreover, convergence to these suboptimal working model elements might occur slowly due to data splitting.

7 Relationship to Classical Methods

It may be useful to compare empirical efficiency maximization to more traditional methods for covariate adjustment in randomized experiments.

7.1 ANCOVA

With a continuous response Y, the well-known ANCOVA technique is to fit the linear regression model

$$Y_i = \alpha + \gamma\Delta_i + \beta^T W_i + \epsilon_i \tag{11}$$

with least squares, and estimate the intention-to-treat parameter with the coefficient estimate γ_n. Under a slightly altered fixed design setting, Freedman (2007a) notes the estimator is consistent and asymptotically Gaussian even when the linear regression model is misspecified, but shows asymptotic variance can be better or worse than when using the unadjusted estimator.

It is simple algebra to show the classical ANCOVA method reduces to fitting $E[Y|\Delta, W]$ with linear regression as in (11), and using the fit to approximate nuisance parameters in locally efficient estimation.

The resulting $Q_n(\cdot)$ applied in parameter estimate μ_{n,Q_n} is a linear function of the covariates. Empirical efficiency maximization could instead attempt to find the optimal linear function $Q(\cdot)$ for use in parameter estimate $\mu_{n,Q}$. The procedure reduces to adding observation weights $|\frac{\Delta_i}{\pi_n} - \frac{1-\Delta_i}{1-\pi_n}|^2$ to the linear regression of the response Y on the covariate vector W, and as previously mentioned leads to asymptotic equivalence with the estimator of Koch et al. (1998) and several estimators pointed to by Tsiatis et al. (2000).

Special linear functions Q are the constant functions $Q(W) = c$, and the limit (in $L^2(P_W)$) of the Q_n corresponding to the ANCOVA estimator. Consequently, the empirical efficiency maximization estimator can only improve upon asymptotic variance relative to these two standard approaches.

When the model (11) is correct in that $E[Y|\Delta, W] = \alpha + \gamma\Delta + \beta^T W$, the ANCOVA and empirical efficiency maximization estimators will be asymptotically efficient.

This efficiency does not contradict a well-known fact sometimes used to argue against covariate adjustment: even when the linear model (11) is correct, with independent Gaussian errors with constant variance, the ANCOVA estimator can lose power relative to the unadjusted estimator for testing treatment effects. If covariates are not predictive of the response, there is little residual variance reduction to improve precision of the treatment coefficient γ, but resulting covariate-adjusted tests lose degrees of freedom. Such considerations will disappear with enough data, as the asymptotic variance of the ANCOVA estimator can never exceed that of the unadjusted estimator when the regression model is correct. However, whether $\sqrt{n}$-asymptotics guide performance in practical problems is determined by sample size, dimensionality of covariate vector, and the true data generating distribution, and empirical efficiency maximization estimators are likely to suffer from second-order error when utilizing unpredictive covariates. Such error may be relevant in clinical trials, for which Pockock et al. (2002) remark "most covariates are not strongly related to the outcome."

7.2 Logistic Regression

Consider a binary response $Y \in \{0, 1\}$, for which the intention-to-treat parameter becomes the difference in probabilities $\mu = P(Y = 1|\Delta = 1) - P(Y = 1|\Delta = 0)$. A logistic regression model specifies

$$\text{logit } P(Y=1|\Delta, W) = \alpha + \gamma\Delta + \beta^T W,$$

and coefficients would typically be fit by maximizing the likelihood with an iterative algorithm. When modeling the covariate distribution $\mathcal{L}(W)$ with the empirical distribution placing mass $\frac{1}{n}$ on each of $W_1, ..., W_n$, one can form a plug-in estimator for the intention to treat parameter $\mu = E[E[Y|\Delta = 1, W]] - E[E[Y|\Delta = 0, W]]$. This takes the form

$$\mu_n = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{1+\exp(-\alpha_n - \gamma_n - \beta_n^T W_i)} - \frac{1}{n}\sum_{i=1}^{n} \frac{1}{1+\exp(-\alpha_n - \beta_n^T W_i)}. \tag{12}$$

Such a plug-in estimator is studied in Moore and van der Laan (2007) and Freedman (2007b), and surprisingly is consistent even if the underlying logistic regression model is misspecified.

Recalling that $Q^\star(W) = (1-\pi_n)E[Y|\Delta = 1, W] + \pi_n E[Y|\Delta = 0, W]$ is the optimal Q for entry into $\mu_{n,Q}$, one could also consider substituting the logistic regression fit of $E[Y|\Delta, W]$ into $Q^\star$ to form Q_n, and then forming parameter estimate μ_{n,Q_n}. In fact, it can be shown that such an estimator is algebraically equivalent to the plug-in (12). Hence, plugging-in for the intention-to-treat parameter based on a logistic regression fit corresponds to standard likelihood-based locally efficient estimation, upon which empirical efficiency maximization attempts to improve.

8 Inference

A strong null hypothesis is that the treatment has no effect, so the counterfactual outcomes Y_T and Y_C are identically distributed. It is common knowledge that one can form exact p-values based on any test statistic $T_n = T_n(O_1, ..., O_n)$. To find the null distribution, repeatedly permute treatment labels and recompute the test statistic. This procedure could of course be applied with our μ_{n,Q_n}.

If inference is desired for the intention-to-treat parameter $\mu = E[Y_T] - E[Y_C]$, we note that if $Q_n(\cdot)$ converges to some $Q_0(\cdot)$ as in Lemma 2, then $\sqrt{n}(\mu_n - \mu) \to N(0, \sigma^2(Q_0, \pi))$. Asymptotically valid inference can be conducted by consistently estimating $\sigma^2(Q_0, \pi)$ with σ_n^2, because by Slutsky's Theorem $\frac{\mu_{n,Q_n} - \mu}{\sigma_n/\sqrt{n}} \to N(0,1)$. Based on preliminary fits $\mu_{T,n}$ and $\mu_{C,n}$, one could form

$$\sigma_n^2 = MSE_n(Q_n, \pi_n) - \frac{1}{\pi_n}|\mu_{T,n} - \frac{1}{n}\sum_{i=1}^{n} Q_n(W_i)|^2 - \frac{1}{1-\pi_n}|\mu_{C,n} - \frac{1}{n}\sum_{i=1}^{n} Q_n(W_i)|^2. \tag{13}$$

The preliminary estimates $\mu_{T,n}$ and $\mu_{C,n}$ will generally only need to be consistent, and unadjusted sample means in the treatment and control groups might be convenient.

Like the parameter estimate itself, the asymptotic variance estimate σ_n^2 might suffer from a finite sample overfitting bias of the weighted mean squared error. The problem vanishes with enough data, but could be acute with relatively small sample sizes or large working models. Our initial recommendation is to perform inference by bootstrapping within the treatment and control groups.

9 Numerical Asymptotic Efficiency Calculations

We assessed estimator performance by generating data according to the following mechanism.

$$\begin{aligned}
&W \sim N(0,1)\\
&\pi_n = \frac{1}{2}\\
&\text{logit } P(Y=1|\Delta, W) = -1 + \Delta + W + \eta W^2.
\end{aligned}$$

That is, covariate W was drawn from a univariate standard normal distribution, half the subjects were assigned treatment, and conditional outcome probabilities in the treatment and control groups grew sigmoidally in slightly different quadratic functions of the covariate. The scalar η was a model misspecification parameter, determining the misspecification of a logistic regression model for the regression of Y on $\{W, \Delta\}$.

We examined four estimators. The first was the unadjusted (4) making no use of covariate information. The second was a likelihood-based locally efficient estimator, which used a fit Q_n of $Q^\star(W) = (1-\pi_n)E[Y_T|W] + \pi_n E[Y_C|W]$ based on separate logistic regression fits for $E[Y_T|W] = E[Y|\Delta = 1, W]$ and $E[Y_C|W] = E[Y|\Delta = 0, W]$ in the treatment and control groups, and then applied μ_{n,Q_n}. We also computed an empirical efficiency maximization estimator, corresponding to positing the working model $Q^\star(W) = \frac{1}{1+\exp(-\alpha-\beta W)}$. As previously discussed, we minimized the weighted nonlinear least squares

$$MSE_n(Q_{c,\alpha,\beta}, \pi_n) = \frac{1}{n}\sum_{i=1}^{n}(\frac{\Delta_i}{\pi_n} - \frac{1-\Delta_i}{1-\pi_n})^2 |Y_i - c - \frac{1}{1+\exp(-\alpha-\beta W)}|^2$$

over (c, α, β) to form (c_n, α_n, β_n), and then applied μ_{n,Q_n} with $Q_n(W) = \frac{1}{1+\exp(-\alpha_n-\beta_n W)}$. The **nlminb()** function in the R language solved the optimization problem, when given starting values of $c = \alpha = \beta = 0$. Finally, we could compute the efficient estimator $\mu_{n,Q^\star}$, because the optimal $Q^\star(W) = \frac{1-\pi_n}{1+\exp(-W-W^2)} + \frac{\pi_n}{1+\exp(1-W-W^2)}$ was known.

Asymptotic variances of the four estimators are displayed in Figure 1, as the model misspecification parameter η grows from 0 to 3 in steps of 0.05. When divided by the asymptotic variance of the efficient estimator, these become asymptotic relative efficiencies, and are shown in Figure 2.

Based on a sample of size $n = 100,000$, we approximated the limiting $Q_n(\cdot)$ corresponding to the likelihood-based locally efficient estimator and empirical efficiency maximization estimator. These were known by design for the unadjusted and efficient estimators. Another sample of the same size allowed computation of μ_T and μ_C. Based on the limit function $Q_0(\cdot)$ for the four estimators, we then used another sample of size $n = 100,000$ to approximate the asymptotic variance $\sigma^2(Q_0, \pi)$ as in (13).

When there was little model misspecification, the three estimators making use of covariate information had roughly equal performance, were close to efficient, and outperformed the unadjusted estimator. As model misspecification increased, the three covariate-adjusted estimators continued outperforming the unadjusted estimator, but empirical efficiency maximization seemed preferable to the standard locally efficient procedure. For model misspecification close to $\eta = 3$, the standard covariate-adjusted procedure appeared equivalent to using unadjusted analysis, while empirical efficiency maximization was superior.

10 Simulations

We can think of at least two objections to these figures. The first is that we haven't yet proven the $Q_n(\cdot)$ of empirical efficiency maximization always converges to some $Q_0(\cdot)$, so we can't strictly assume $\sqrt{n}(\mu_{n,Q_n} - \mu)$ is asymptotically Gaussian, with an asymptotic variance. Another criticism could be that our results might not relay performance for realistic sample sizes.

We thus supplemented our Monte Carlo asymptotic variance calculations with simulations. For each model misspecification parameter η, we generated $10,000$ datasets of size $n = 100$. For each dataset, we formed the four estimators μ_{n,Q_n}. By averaging over the $10,000$ datasets, we were able to approximate the error $E|\mu_{n,Q_n} - \mu|^2$, where the intention-to-treat parameter μ was found in our previous Monte Carlo analysis. The unadjusted and efficient estimators were also exactly unbiased, while the likelihood-based locally efficient and empirical efficiency maximization estimators were not. In particular, the unbiased $\mu_{n,Q^\star}$ had the smallest variance among any $\mu_{n,Q}$, not only the smallest asymptotic variance. Figure 3 displays the estimators' mean squared errors, when standardized by the mean squared error $E|\mu_{n,Q^\star} - \mu|^2$ of the efficient estimator. There appears to be a small amount of simulation error, but results clearly demonstrate the previous asymptotic calculations are informative for realistic sample sizes, because Figure 3 essentially replicates Figure 2.

11 Discussion

We have proposed a new method for utilizing baseline covariate information when estimating the intention-to-treat parameter in randomized experiments. The method is quite general, yet can be simply stated and easily implemented. With binary outcomes, our procedure is highly dissimilar to existing techniques.

To summarize, we propose building a working model for prediction of response Y from covariates W, and fitting the model to minimize a weighted mean squared error. Observations in the treatment and control groups are given weights proportional to π_n^{-2} and $(1 - \pi_n)^{-2}$, where π_n is the proportion of subjects assigned to treatment. The resulting function Q_n is then entered into μ_{n,Q_n} of (5) to form the parameter estimate.

Consistency and asymptotic Gaussianity are guaranteed under the minimal conditions of Lemma 2, and are not compromised through utilizing misspecified working models. Instead of fitting working models to maximize likelihood, we improve upon standard locally efficient estimation by aiming at the working model element minimizing asymptotic variance for the resulting intention-to-treat estimator.

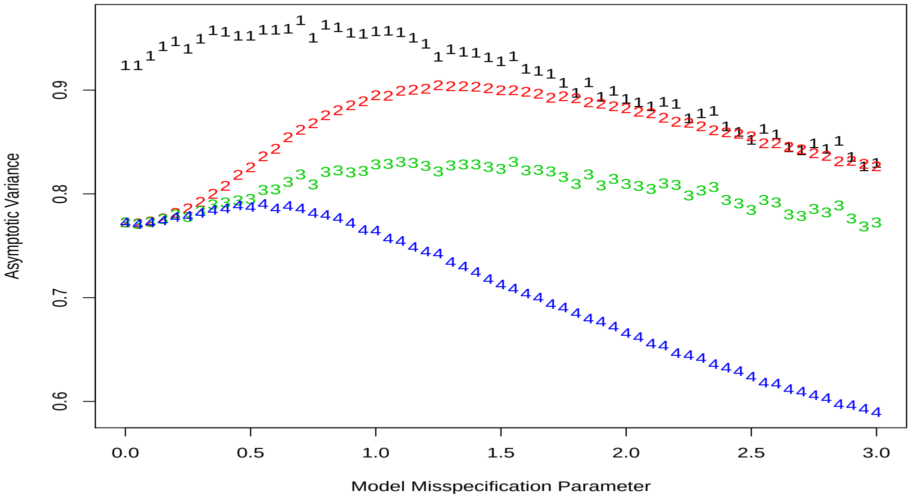

Figure 1: 1 − 4 represent asymptotic variances of the unadjusted estimator, locally efficient estimator fitting a misspecified logistic regression model in each stratum, empirical efficiency maximization estimator fitting a misspecified logistic regression model with weighted nonlinear least squares, and the efficient estimator.

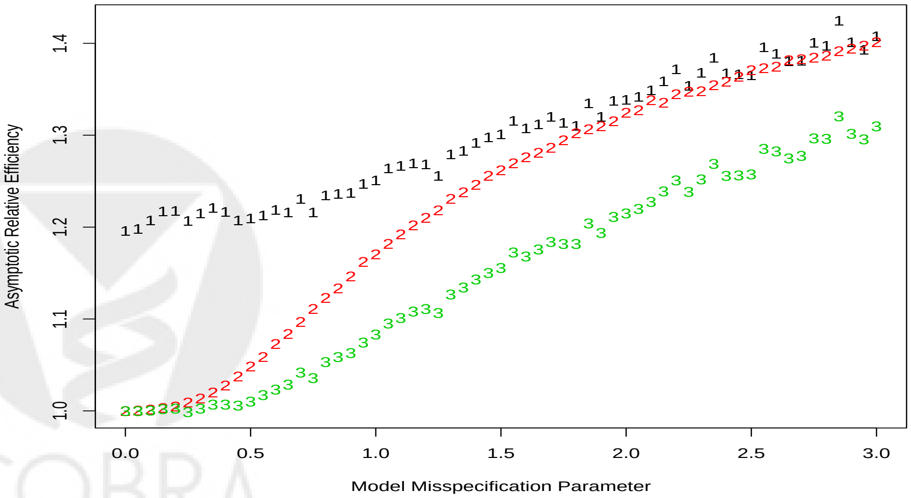

Figure 2: 1 − 3 represent the asymptotic variances of the unadjusted, likelihood-based locally efficient, and empirical efficiency maximization estimators, when divided by the asymptotic variance of the efficient estimator.

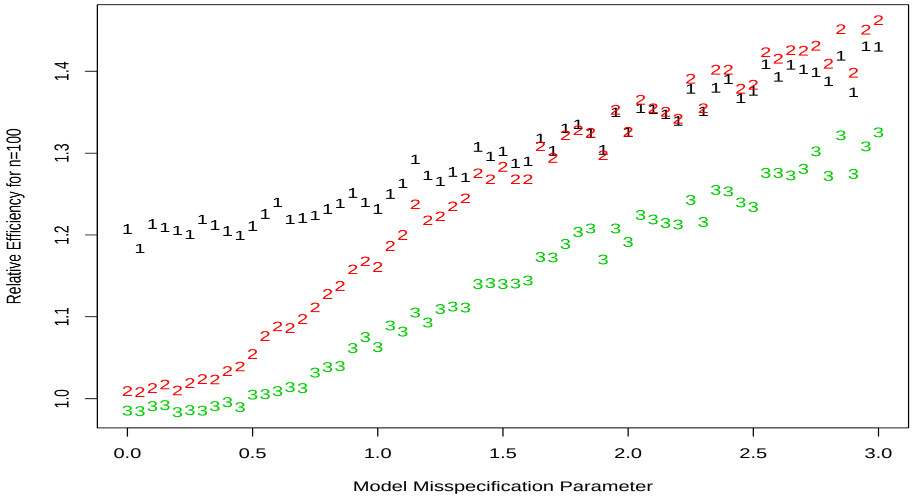

Figure 3: For a sample size of $n = 100$, labels $1-3$ represent the ratio of the mean squared error of the unadjusted, likelihood-based locally efficient, and empirical efficiency maximization estimators to the mean squared error of the efficient estimator.

Appendix: Proofs of Lemmas 1-3

Remark on proofs. Consider two i.i.d. samples $\{W_i, Y_{T,i}\}_{i=1}^{m_n}$ and $\{W_j', Y_{C,j}'\}_{j=1}^{n-m_n}$, which are independent of each other. Estimation of $\mu = \mu_T - \mu_C$ in this setting is a special case of the two-sample problem, where covariates are observed for each subject, but have the same marginal distribution in the two populations. Form the estimator

$$\hat{\mu}_{n,Q} = \frac{1}{n}\{\sum_{i=1}^{m_n} \frac{1}{\pi_n}(Y_{T,i} - Q(W_i)) - \sum_{j=1}^{n-m_n} \frac{1}{1-\pi_n}(Y_{C,j}' - Q(W_j')\}.$$

This clearly has the same distribution as $\mu_{n,Q} = \frac{1}{n}\sum_{i=1}^{n}(\frac{\Delta_i}{\pi_n} - \frac{1-\Delta_i}{1-\pi_n})(Y_i - Q(W_i))$ under consideration. Hence, it suffices to analyze $\hat{\mu}_{n,Q}$ under the altered sampling scheme.

When $Q_n(\cdot;\ O_1, ..., O_n)$ does not depend on the ordering of $O_1, ...O_n$ (as assumed in Lemma 2), we can likewise note the distribution of μ_{n,Q_n} is equal to that of $\hat{\mu}_{n,\hat{Q}_n}$, where $\hat{Q}_n(\cdot) = Q_n(\cdot;\ (W_i, 1, Y_{T,i}), (W_j', 0, Y_{C,j}'),\ i = 1, ..., m_n,\ j = 1, ..., n - m_n)$.

In the proofs of Lemma 1 and Lemma 2, we will thus analyze estimators in the two-sample setting. When assumed conditions on $\mu_{n,Q}$ or Q_n correspond naturally to conditions on $\hat{\mu}_{n,Q}$ or $\hat{Q}_n$, these will be invoked without explicit mention.

Proof of Lemma 1. Simple algebra yields that

$$\begin{aligned}\sqrt{n}(\hat{\mu}_{n,Q} - \mu) &= \pi_n^{-1/2} m_n^{-1/2} \sum_{i=1}^{m_n}(Y_{T,i} - Q(W_i) - \mu_T - E[Q(W)]) \\ &- (1-\pi_n)^{-1/2}(n-m_n)^{-1/2} \sum_{j=1}^{n-m_n}(Y_{C,j}' - Q(W_j') - \mu_C - E[Q(W)]).\end{aligned}$$

The unbiasedness of $\hat{\mu}_{n,Q}$ follows immediately, as all terms in each of the two sums have mean zero. Asymptotic normality also follows from the independence of the two samples and the Central Limit Theorem, and it remains to evaluate the variance. As the above expression is a sum of independent terms, the variance can be computed as

$$\begin{aligned}
\sigma^2(Q,\pi_n) &= \frac{1}{\pi_n}\mathrm{Var}(Y_T - Q(W)) + \frac{1}{1-\pi_n}\mathrm{Var}(Y_C - Q(W)). \\
&= \frac{1}{\pi_n}E|Y_T - Q(W)|^2 - (\mu_T - E[Q(W)])^2 \\
&+ \frac{1}{1-\pi_n}E|Y_C - Q(W)|^2 - (\mu_C - E[Q(W)])^2, \qquad (14)
\end{aligned}$$

where we express variance as mean squared error subtracting off squared bias.

To represent this variance, we now depart from the two-sample setting and return to the sampling scheme considered in this paper's body. Because $X = (W, Y_T, Y_C) \perp \Delta$ by the randomization, we note that for any integrable $\psi(W,Y)$,

$$\begin{aligned}
E[\frac{\Delta}{\pi_n}\psi(W,Y)] &= E[\frac{\Delta}{\pi_n}\psi(W,Y_T)] = E[\frac{\psi(W,Y_T)}{\pi_n}E[\Delta|W,Y_T]] \\
&= E[\frac{\psi(W,Y_T)}{\pi_n}E[\Delta]] = E[\frac{\psi(W,Y_T)}{\pi_n}\pi_n] = E[\psi(W,Y_T)],
\end{aligned}$$

and that likewise $E[\frac{1-\Delta}{1-\pi_n}\psi(W,Y)] = E[\psi(W,Y_C)]$. Hence,

$$\begin{aligned}
E|Y_T - Q(W)|^2 &= E[\frac{\Delta}{\pi_n}|Y - Q(W)|^2] \\
E|Y_C - Q(W)|^2 &= E[\frac{1-\Delta}{1-\pi_n}|Y - Q(W)|^2],
\end{aligned}$$

so (14) tells us

$$\begin{aligned}
&\sigma^2(Q,\pi_n) - \frac{1}{\pi_n}(\mu_T - E[Q(W)])^2 - \frac{1}{1-\pi_n}(\mu_C - E[Q(W)])^2 \\
=\ & E[(\frac{\Delta}{\pi_n^2}|Y - Q(W)|^2] + E[\frac{1-\Delta}{(1-\pi_n)^2}|Y - Q(W)|^2 \\
=\ & E[(\frac{\Delta}{\pi_n^2} + \frac{1-\Delta}{(1-\pi_n)^2})|Y - Q(W)|^2] = E[(\frac{\Delta}{\pi_n} - \frac{1-\Delta}{1-\pi_n})^2|Y - Q(W)|^2],
\end{aligned}$$

or the variance result given in the lemma. □

Proof of Lemma 2. Let $P_{T,n}$ denote the empirical distribution placing mass $\frac{1}{m_n}$ on each of $(W_1, Y_{T,1}), ..., (W_{m_n}, Y_{T,m_n})$, and let $P_{C,n}$ denote the empirical distribution placing mass $\frac{1}{n-m_n}$ on each of $(W_1', Y_{C,1}'), ..., (W_{n-m_n}', Y_{C,n-m_n}')$. Note that covariate W

has a common marginal distribution P_W in the two populations. We write

$$\begin{aligned}
\hat{\mu}_{n,\hat{Q}_n} - \hat{\mu}_{n,Q_0} &= \{\int (y_T - \hat{Q}_n(w))dP_{T,n}(w, y_T) - \int (y_C - \hat{Q}_n(w))dP_{C,n}(w, y_C)\} \\
&- \{\int (y_T - Q_0(w))dP_{T,n}(w, y_T) - \int (y_C - Q_0(w))dP_{C,n}(w, y_C)\} \\
&= \int (Q_0(w) - \hat{Q}_n(w))dP_{T,n}(w) - \int (Q_0(w) - \hat{Q}_n(w))dP_{C,n}(w) \\
&= \int (Q_0 - \hat{Q}_n)(dP_{T,n} - dP_W) - \int (Q_0 - \hat{Q}_n)(dP_{C,n} - dP_W) \\
&+ \int (Q_0 - \hat{Q}_n)dP_W - \int (Q_0 - \hat{Q}_n)dP_W \\
&= \int (Q_0 - \hat{Q}_n)(dP_{T,n} - dP_W) - \int (Q_0 - \hat{Q}_n)(dP_{C,n} - dP_W).
\end{aligned}$$

Define the seminorm $\rho(Q) = \sqrt{E|Q(W) - E[Q(W)]|^2}$, which is simply the standard deviation of $Q(W)$. Because $\mathcal{Q}_0$ is assumed to be a Donsker class, the asymptotic continuity condition (2.1.8) of van der Vaart and Wellner (1996) tells us

$$\begin{aligned}
0 &= \lim_{\delta \downarrow 0} \limsup_{m_n \to \infty} P^\star(\sup_{\{Q:\ \rho(Q-Q_0)<\delta\}} m_n^{1/2} | \int (Q - Q_0)(dP_{T,n} - dP_W)| > \epsilon) \\
&= \lim_{\delta \downarrow 0} \limsup_{n - m_n \to \infty} P^\star(\sup_{\{Q:\ \rho(Q,Q_0)<\delta\}} (n - m_n)^{1/2} | \int (Q - Q_0)(dP_{C,n} - dP_W)| > \epsilon).
\end{aligned}$$

.

Here $P^\star$ refers to outer probability, to handle any issues involving measurability. For a sequence $\{Q_k\}$ in $\mathcal{Q}_0$, note that the $L^2(P_W)$ convergence $d(Q_k, Q_0) \to 0$ implies the convergence $\rho(Q_k, Q_0) \to 0$. Because $d(\hat{Q}_n, Q_0)$ converges to zero in probability by assumption, this reveals $\int (Q_0 - \hat{Q}_n)(dP_{T,n} - dP_W)$ and $\int (Q_0 - \hat{Q}_n)(dP_{C,n} - dP_W)$ are $o_P(n^{-1/2})$. Slutsky's theorem then tells us $\sqrt{n}(\hat{\mu}_{n,\hat{Q}_n} - \mu)$ and $\sqrt{n}(\hat{\mu}_{n,Q_0} - \mu)$ have the same limiting distribution, which is $N(0, \sigma^2(Q, \pi))$ by Lemma 1, proving the result. □

Proof of Lemma 3A. Note that $\frac{1}{n}\sum_{i=1}^n \Delta_i = \pi_n$, so $\frac{1}{n}\sum_{i=1}^n (\frac{\Delta_i}{\pi_n} - \frac{1-\Delta_i}{1-\pi_n}) = 0$. Hence, $\mu_{n,Q} = \mu_{n,Q+c}$ for any constant c. The two estimators then of course have the same variance, and asymptotic variance.

Proof of Lemma 3B. We next observe that

$$MSE(Q+c, \pi_n) - \sigma^2(Q+c, \pi_n) = \frac{1}{\pi_n}|\mu_T - E[Q(W)] - c|^2 - \frac{1}{1-\pi_n}|\mu_C - E[Q(W)] - c|^2.$$

The convex quadratic in c is minimized at $c^\star = (1 - \pi_n)\mu_T + \pi_n\mu_C - E[Q(W)]$. As $\sigma^2(Q + c, \pi_n)$ doesn't depend on c, this $c^\star$ must minimize $MSE(Q + c, \pi_n)$.

Proof of Lemma 3C. Suppose $MSE(Q, \pi_n) \le MSE((1 - \pi_n)\mu_T + \pi_n\mu_C, \pi_n)$. By (A), $\sigma^2(Q, \pi_n) = \sigma^2(Q + (1 - \pi_n)\mu_T + \pi_n\mu_C - E[Q(W)], \pi_n)$. Note also from examining

$\sigma^2(Q,\pi_n) - MSE(Q,\pi_n)$ that

$$\begin{aligned}&\sigma^2(Q + (1-\pi_n)\mu_T + \pi_n\mu_C - E[Q(W)],\pi_n) - \sigma^2((1-\pi_n)\mu_T + \pi_n\mu_C,\pi_n)\\ &= MSE(Q + (1-\pi_n)\mu_T + \pi_n\mu_C - E[Q(W)],\pi_n) - MSE((1-\pi_n)\mu_T + \pi_n\mu_C,\pi_n),\end{aligned}$$

because both $Q(W)+(1-\pi_n)\mu_T+\pi_n\mu_C-E[Q(W)]$ and the constant $(1-\pi_n)\mu_T+\pi_n\mu_C$ have the same expectation. Hence,

$$\begin{aligned}&\sigma^2(Q,\pi_n) - \sigma^2((1-\pi_n)\mu_T + \pi_n\mu_C,\pi_n)\\ &= \sigma^2(Q + (1-\pi_n)\mu_T + \pi_n\mu_C - E[Q(W)],\pi_n) - \sigma^2((1-\pi_n)\mu_T + \pi_n\mu_C,\pi_n)\\ &= MSE(Q + (1-\pi_n)\mu_T + \pi_n\mu_C - E[Q(W)],\pi_n) - MSE((1-\pi_n)\mu_T + \pi_n\mu_C,\pi_n)\\ &\leq MSE(Q,\pi_n) - MSE((1-\pi_n)\mu_T + \pi_n\mu_C,\pi_n) \leq 0,\end{aligned}$$

Proof of Lemma 3D. We can by (B) without loss of generality take $Q_1 = Q_0 + c^\star$, for $Q_0 \in \mathcal{Q}$ and $c^\star = (1-\pi_n)\mu_T + \pi_n\mu_C - E[Q_0(W)]$. Then for any $Q \in \mathcal{Q}$, we observe that $MSE(Q_1,\pi_n) \leq MSE(Q + (1-\pi_n)\mu_T + \pi_n\mu_C - E[Q(W)],\pi_n)$ and (A) imply

$$\begin{aligned}\sigma^2(Q_0,\pi_n) - \sigma^2(Q,\pi_n)&\\ = \sigma^2(Q_0 + c^\star,\pi_n) - \sigma^2(Q + (1-\pi_n)\mu_T + \pi_n\mu_C - E[Q(W)],\pi_n)&\\ = MSE(Q_0 + c^\star,\pi_n) - MSE(Q + (1-\pi_n)\mu_T + \pi_n\mu_C - E[Q(W)],\pi_n)&\\ = MSE(Q_1,\pi_n) - MSE(Q + (1-\pi_n)\mu_T + \pi_n\mu_C - E[Q(W)],\pi_n) \leq 0.&\end{aligned}$$

Proof of Lemma 3E. As the class of all square integrable functions from Support(W) to $\mathbb{R}$ is clearly closed under shifts, the just proven (D) tells us it suffices to consider the function minimizing $MSE(Q,\pi_n)$. From (14) derived in the proof of Lemma 1, we have that

$$\begin{aligned}MSE(Q,\pi_n) = \frac{1}{\pi_n}E|Y_T - Q(W)|^2 + \frac{1}{1-\pi_n}E|Y_C - Q(W)|^2&\\ = \frac{1}{\pi_n}E[Y_T^2 - 2Y_TQ(W) + Q^2(W)] + \frac{1}{1-\pi_n}E[Y_C^2 - 2Y_CQ(W) + Q^2(W)]&\\ = E[\frac{1}{\pi_n}\{Y_T^2 - 2E[Y_T|W]Q(W) + Q^2(W)\} + \frac{1}{1-\pi_n}\{Y_C^2 - 2E[Y_C|W]Q(W) + Q^2(W)\}],&\end{aligned}$$

where the last line follows from conditioning on W. Within the expectation, the convex quadratic in $Q(W)$ is minimized at the given $Q^\star(W)$, concluding the proof. □

References

[1] Fisher, R.A. (1932). *Statistical Methods for Research Workers.* Edinburgh: Oliver and Boyd (13th ed., 1958).

[2] Freedman, D.A. (2007a). On regression adjustments to experimental data. To appear in *Advances in Applied Mathematics.*

[3] Freedman, D.A. (2007b). Randomization does not justify logistic regression. Available at http://www.stat.berkeley.edu/ census/neylogit.pdf

[4] Koch GG, Tangen CM, Jung JW, Amara IA (1998). Issues for covariance analysis of dichotomous and ordered categorical data from randomized clinical trials and nonparametric strategies for addressing them. *Statistics in Medicine*, 17, 1863-1892.

[5] Moore, K.L. and van der Laan, M.J. (2007). Covariate adjustment in randomized trials with binary outcomes: Targeted maximum likelihood estimation. *U.C. Berkeley Division of Biostatistics Working Paper Series*. Working paper 215.

[6] Neyman, J. (1923) Sur les applications de la théorie des probabilitiés aux experiences agricoles: Essai des principes. *Roczniki Nauk Rolniczych*, 10, 1-51, in Polish. English translation by D.M. Dabrowska and T.P. Speed (1990), *Statistical Science*, 5, 465-480 (with discussion).

[7] Pocock, S.J., Assmann, S.E., Enos, L.E., and Kasten, L.E. (2002). Subgroup analysis, covariate adjustment, and baseline comparisons in clinical trial reporting: current practice and problems. *Statistics in Medicine*, 21, 2917-2930.

[8] Robins, J.M. and Rotnitzky, A. (1992). Recovery of information and adjustment for dependent censoring using surrogate markers. *AIDS Methodology - Methodological Issues.* Jewell, N., Dietz, K, and Farewell, W., eds. Birkhäuser, Boston, 297-331.

[9] Robins, J.M., Rotnitzky, A., and Zhao, L.P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89, 846-866.

[10] Rubin, D.B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66, 688-701.

[11] Rubin, D.B. and van der Laan, M.J. (2008). Empirical efficiency maximization. *International Journal of Biostatistics*, in press. Preliminary technical report from 2007 available at http://www.bepress.com/ucbbiostat/paper220.

[12] Sontag, E.D. and Sussmann, H.J. (1989). Backpropagation can give rise to spurious local minima even for networks without hidden layers. *Complex Systems*, 3, 91-106.

[13] Tsiatis, A.A. (2006). *Semiparametric Theory and Missing Data*. Springer Science + Business Media, LLC.

[14] Tsiatis, A.A., Davidian, M., Zhang, M., and Lu, X. (2000). Covariate adjustment for two-sample treatment comparisons in randomized clinical trials: A principled yet flexible approach. *Statistics in Medicine*, 25, 1-10. Revised December 30, 2006.

[15] van der Laan, M.J. and Robins, J.M. (2003). *Unified Methods for Censored Longitudinal Data and Causality*. Springer-Verlag, New York.

[16] van der Vaart, A.W. and Wellner, J.A. (1996). *Weak Convergence and Empirical Processes with Applications to Statistics*. Springer-Verlag, New York.