The International Journal of Biostatistics

Volume 4, Issue 1

2008

Article 5

Empirical Efficiency Maximization: Improved Locally Efficient Covariate Adjustment in Randomized Experiments and Survival Analysis

Daniel B. Rubin, *University of California, Berkeley*

Mark J. van der Laan, *University of California, Berkeley*

Recommended Citation:

Rubin, Daniel B. and van der Laan, Mark J. (2008) "Empirical Efficiency Maximization: Improved Locally Efficient Covariate Adjustment in Randomized Experiments and Survival Analysis," *The International Journal of Biostatistics*: Vol. 4: Iss. 1, Article 5.

DOI: 10.2202/1557-4679.1084

Empirical Efficiency Maximization: Improved Locally Efficient Covariate Adjustment in Randomized Experiments and Survival Analysis

Daniel B. Rubin and Mark J. van der Laan

Abstract

It has long been recognized that covariate adjustment can increase precision in randomized experiments, even when it is not strictly necessary. Adjustment is often straightforward when a discrete covariate partitions the sample into a handful of strata, but becomes more involved with even a single continuous covariate such as age. As randomized experiments remain a gold standard for scientific inquiry, and the information age facilitates a massive collection of baseline information, the longstanding problem of if and how to adjust for covariates is likely to engage investigators for the foreseeable future.

In the locally efficient estimation approach introduced for general coarsened data structures by James Robins and collaborators, one first fits a relatively small working model, often with maximum likelihood, giving a nuisance parameter fit in an estimating equation for the parameter of interest. The usual advertisement is that the estimator will be asymptotically efficient if the working model is correct, but otherwise will still be consistent and asymptotically Gaussian.

However, by applying standard likelihood-based fits to misspecified working models in covariate adjustment problems, one can poorly estimate the parameter of interest. We propose a new method, empirical efficiency maximization, to optimize the working model fit for the resulting parameter estimate.

In addition to the randomized experiment setting, we show how our covariate adjustment procedure can be used in survival analysis applications. Numerical asymptotic efficiency calculations demonstrate gains relative to standard locally efficient estimators.

KEYWORDS: empirical efficiency maximization, covariate adjustment, locally efficient estimation, two-phase designs, clinical trials, survival analysis

1 Introduction

Consider an experiment in which covariate vectors $\{W_i\}_{i=1}^n$ are measured on n subjects at baseline, the subjects are randomly assigned to treatment or control arms, and outcomes $\{Y_i\}_{i=1}^n$ are then assessed. The observed data could be written as $\{W_i, \Delta_i, Y_i\}_{i=1}^n$, for $\Delta_i = 1$ and $\Delta_i = 0$ indicating assignment to treatment or control. Suppose interest lies in the *intention-to-treat parameter*

$$\mu_{ITT} = E[Y|\Delta = 1] - E[Y|\Delta = 0], \quad (1)$$

or difference in expected responses among those assigned to the two groups.

A perfectly valid estimator of (1) could be formed by ignoring baseline covariates $\{W_i\}_{i=1}^n$, and applying

$$\hat{\mu}_{ITT} = \frac{\sum_{i=1}^n \Delta_i Y_i}{\sum_{i=1}^n \Delta_i} - \frac{\sum_{i=1}^n (1 - \Delta_i) Y_i}{\sum_{i=1}^n 1 - \Delta_i}, \quad (2)$$

which is simply the difference of sample means in treatment and control arms.

However, many authors following Fisher (1932) have observed that discarding covariate information is potentially wasteful. The intuition is that a subject's covariate W_i might inform how he or she would have responded in both study arms, while unadjusted analysis cannot exploit such knowledge.

Pocock et al. (2002) recently 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.” As randomized experiments represent a gold standard for scientific inquiry, and the information age facilitates massive collection of covariate information, devising and promulgating “appropriate” methods for adjustment appears to be an important task.

The semiparametric literature (e.g. Bickel et al., 1998) has traditionally focused on asymptotically efficient estimators. Unfortunately, such estimators suffer from the curse of dimensionality in nontrivial covariate adjustment problems, and the asymptotically efficient estimator of the intention-to-treat parameter μ_{ITT} would involve consistently estimating the treatment and control regression functions $Q_T^*(w) = E[Y|W = w, \Delta = 1]$ and $Q_C^*(w) = E[Y|W = w, \Delta = 0]$, and using the fits in an estimating equation. While this could technically be carried out under minimal assumptions, the nonparametric function approximation problem could be very difficult for covariates W of even moderate dimension, and an asymptotically efficient estimator would most likely perform poorly in practice.

A locally efficient estimator is a middle course between inefficient unadjusted estimators and impractical efficient estimators. Here one would fit a relatively small working model to approximate the unknown regression functions with $\hat{Q}_T$ and $\hat{Q}_C$, and use them as nuisance parameter fits in the efficient estimating equation for the intention-to-treat parameter. Under general conditions, asymptotic efficiency can be achieved if the working model holds, while a misspecified working model will not compromise $\sqrt{n}$ -consistency or asymptotic Gaussianity. The goal is to gain precision by making some use of informative covariates, while controlling stability by restricting the working model's size.

The way locally efficient estimation has been presented for coarsened data structures by Robins and Rotnitzky (1992), Robins, Rotnitzky, and Zhao (1994), van der Laan and Robins (2003), Tsiatis (2006), and others, the working model fit would most often be identical to the fit of someone who believed the model actually held. For instance, if using a logistic regression model for the binary regression of outcome Y on (W, Δ) , coefficients would be fit with maximum likelihood, and implemented with iteratively re-weighted least squares. When the working model is incorrect, such a fit can be a very poor choice for the resulting treatment effect estimate. In fact, performance can degrade relative to the unadjusted estimator making no use of covariate information. In this work we present a new locally efficient technique, *empirical efficiency maximization*, which aims to select the optimal working model element for estimating the parameter of interest, irrespective of whether the working model is correctly specified.

We introduce our method in the following section, and request the reader's indulgence as we motivate it in an abstract setting. For anyone disinclined to peruse our general formulation, but interested in our proposals for using working models for covariate adjustment, little will be lost by skipping ahead. We return to randomized experiments in Section 3, and survival analysis applications are examined in Section 4. Numerical asymptotic efficiency calculations demonstrate advantages relative to existing methods.

2 General Coarsened Data Formulation

Suppose that in an ideal world we would take an i.i.d. sample $\{X_i\}_{i=1}^n$, where $X \sim F_0 \in \mathcal{F}$ contains a subject's full data. Here F_0 is the unknown data generating distribution, and $\mathcal{F}$ is a statistical model. Suppose our interest would be in estimating a full data parameter $\mu = \mu(F_0) \in \mathbb{R}$. For the time being, we will restrict attention to estimating a population mean $\mu = E_{F_0}[\psi(X)] \in \mathbb{R}$, and defer treatment of more general parameters to Section 5.

If we could observe the full data, we could estimate μ with the empirical mean $\frac{1}{n} \sum_{i=1}^n \psi(X_i)$. But due to missingness, censoring, or other problems, we often aren't able to measure everything we'd like to about each subject. Hence, assume we only have access to a coarsened dataset $\{O_i\}_{i=1}^n$, where

$$O = \Phi(X, C) \sim P_0 = P_{F_0, G_0} \in \mathcal{M} = \{P_{F, G} : F \in \mathcal{F}, G \in \mathcal{G}\}.$$

The Φ will be a known function, and the coarsening variable C will determine how much of X is actually observed. The data generating distribution for O is P_0 , belonging to statistical model $\mathcal{M}$. The law G_0 of $\{C|X\}$ is known as the coarsening mechanism, belonging to the model $\mathcal{G}$. We'll assume the model $\mathcal{G}$ obeys missingness at random as introduced in Heitjan and Rubin (1991), or more generally the coarsening at random of Gill et al. (1997), meaning the probability of missingness or coarsening only depends on a part of the full data that we can always observe.

This paper deals with how to estimate parameter μ when coarsening mechanism G_0 is either known or can be easily estimated, which we'll argue can happen in a variety of real-world examples. To be more specific, we assume we can correctly specify that $G_0 \in \mathcal{G}_0 \subset \mathcal{G}$, where $\mathcal{G}_0$ is a submodel for the coarsening mechanism, and that from this submodel we can efficiently estimate G_0 with a $\hat{G}$.

An estimator sequence $\hat{\mu} = \hat{\mu}(O_1, \dots, O_n)$ is said to be asymptotically linear with influence curve $IC(O|P_0)$ if

$$\hat{\mu} = \mu + \frac{1}{n} \sum_{i=1}^n IC(O_i|P_0) + o_{P_0}(n^{-1/2}), \quad (3)$$

and $IC(O|P_0)$ has mean zero and finite variance. For a full data population mean $\mu = E_{F_0}[\psi(X)]$, an asymptotically efficient estimator's influence curve takes the form

$$IC(O|P_0) = D(O|G_0, Q(F_0)) - \mu \quad (4)$$

when the full data model $\mathcal{F}$ contains no special parametric or semiparametric assumptions, and $\psi(X)$ has finite variance.

Here $D(O|G_0, Q(F_0))$ results from the doubly robust mapping of $\psi(X)$, defined in Theorem 2.1 of van der Laan and Robins (2003). At $P = P_{F, G} \in \mathcal{M}$, the score operator $A : L^2(F) \rightarrow L^2(P_{F, G})$ is given by $A(s)(O) = E_P[s(X)|O]$. Its adjoint $A^T : L^2(P) \rightarrow L^2(F)$ is then $A^T(v)(X) = E_P[v(O)|X]$. Here $L^2(F)$ and $L^2(P)$ respectively refer to the Hilbert spaces of functions of full data X and observed data O that are square integrable under F and P . Defining the information operator as the composition $I = A^T \circ A$, the doubly robust

mapping from $L^2(F)$ to $L^2(P)$ is given by $A \circ I^{-1}$. When evaluated at $\psi(\cdot) \in L^2(F)$, we denote the resulting function in $L^2(P_{F,G})$ by $D(\cdot|G, F)$.

An estimator $\hat{\mu}$ satisfying (3) will be asymptotically Gaussian, meaning $\sqrt{n}(\hat{\mu} - \mu) \rightarrow_{\mathcal{L}} N(0, \sigma^2)$. Asymptotic efficiency at P_0 implies that if there is another regular estimator sequence $\{\hat{\mu}_n\}$ for which $\sqrt{n}(\hat{\mu}_n - \mu) \rightarrow_{\mathcal{L}} N(0, \tau^2)$, then $\sigma^2 \leq \tau^2$, so the sequence estimates parameter μ with less precision. The asymptotically efficient estimator's influence curve is unique, and is termed the efficient influence curve. If $\hat{\mu} - \hat{\mu} = o_{P_0}(n^{-1/2})$, the estimators $\hat{\mu}$ and $\hat{\mu}$ are said to be asymptotically equivalent.

Equations (3) and (4) suggest a route to efficient estimation: fit nuisance parameter $Q(F_0)$ from the data with $\hat{Q}$, and apply $\hat{\mu} = \frac{1}{n} \sum_{i=1}^n D(O_i|\hat{G}, \hat{Q})$. In fact, this can be carried out in many circumstances.

But as we alluded to in the abstract and introduction, and will show in examples, consistently estimating the nuisance parameter $Q(F_0)$ often requires solving a high dimensional function approximation problem, and efficient estimators can be quite unreliable in practice. The locally efficient approach is to instead assume a relatively small working model $\mathcal{F}_0 \subset \mathcal{F}$ for the full data distribution, which induces a working index set

$$\mathcal{Q} = \{Q(F) : F \in \mathcal{F}_0\} \subset \{Q(F) : F \in \mathcal{F}\}.$$

Locally efficient coarsened data estimators have operated by letting $\hat{Q}$ be the efficient estimate of nuisance parameter $Q(F_0)$ under working model $\mathcal{F}_0$. Under regularity conditions, the estimator $\frac{1}{n} \sum_{i=1}^n D(O_i|\hat{G}, \hat{Q})$ will be asymptotically efficient if the working model is correctly specified, meaning $F_0 \in \mathcal{F}_0$, but otherwise will still be consistent and asymptotically linear. This is due to the robustness result in Theorem 2.1 of van der Laan and Robins (2003), that

$$E_{P_0}[D(O|G_0, Q(F))] = \mu(F_0) \text{ for any } F \in \mathcal{F}, \quad (5)$$

meaning a misspecified nuisance parameter $Q(F) \neq Q(F_0)$ will not compromise the estimator.

The problem with local efficiency is that while much can be known about the coarsening mechanism G_0 , the full data working model $\mathcal{F}_0$ will most likely be misspecified in practice. Why trust that a procedure is desirable at our true location F_0 just because it is optimal at other locales? The working set $\mathcal{Q}$ is often simply an index of estimators, excluding many elements of $\{Q(F) : F \in \mathcal{F}\}$ to ease the curse of dimensionality. It is then not necessarily true that standard fits of $\mathcal{F}_0$ are desirable for estimating the parameter of interest.

One way to gauge the quality of $Q \in \mathcal{Q}$ is with the asymptotic variance of $\hat{\mu} = \frac{1}{n} \sum_{i=1}^n D(O_i|G_0, Q)$, which by the robustness result (5) is given by

$$\sigma^2(Q) = \text{Var}_{P_0}(D(O|G_0, Q)) = E_{P_0}[D^2(O|G_0, Q)] - \mu^2. \quad (6)$$

When the full data working model fit $\hat{Q}$ converges to an element $Q \in \mathcal{Q}$, the asymptotic variance of our estimator $\frac{1}{n} \sum_{i=1}^n D(O_i|G_0, \hat{Q})$ will be $\sigma^2(Q)$, and hence it is vital to consider our fit's limiting behavior.

Note that $\hat{\mu}$ is unbiased for the parameter μ , but can only be applied as an estimator with known coarsening mechanism. When our estimator is $\frac{1}{n} \sum_{i=1}^n D(O_i|\hat{G}, Q)$, and $\hat{G}$ is an efficient estimator of G_0 in a correctly specified coarsening mechanism model $\mathcal{G}_0 \subset \mathcal{G}$, then $\sigma^2(Q)$ will actually be an upper bound for our estimator's asymptotic variance. That is, estimating a known coarsening mechanism improves performance. The general phenomenon is discussed by Robins, Rotnitzky, and Zhao (1994), and formal results are stated in Theorem 2.3 of van der Laan and Robins (2003).

In the present work, we consider estimators that attempt to directly find the $Q \in \mathcal{Q}$ minimizing the asymptotic variance bound $\sigma^2(Q)$, or equivalently maximizing a bound for asymptotic efficiency (relative to the asymptotically efficient estimator). The key principle is that $\sigma^2(Q)$ is monotone in the population mean $E_{P_0}[D^2(O|G_0, Q)]$, and we can estimate this population mean with the empirical mean $\frac{1}{n} \sum_{i=1}^n D^2(O_i|\hat{G}, Q)$. The approximation should be valid for all $Q \in \mathcal{Q}$, irrespective of whether this working index set is induced by a correctly specified working model for the data generating distribution. If $\mathcal{Q}$ is not too large a set in the empirical process sense, we might expect the empirical and population means of $D^2(O|G_0, Q)$ to be uniformly close to each other over $\mathcal{Q}$, and for the empirical minimizer to approach the population minimizer. We therefore propose selecting

$$\hat{Q} = \operatorname{argmin}_{Q \in \mathcal{Q}} \frac{1}{n} \sum_{i=1}^n D^2(O_i|\hat{G}, Q),$$

and applying the estimator $\hat{\mu} = \frac{1}{n} \sum_{i=1}^n D(O_i|\hat{G}, \hat{Q})$.

There can in fact be several minimizers of $\sigma^2(Q)$ over working index set $\mathcal{Q}$, and what we precisely mean is finding the function in

$$\mathcal{D} = \{D(\cdot|G_0, Q) : Q \in \mathcal{Q}\} = \{D(\cdot|G_0, Q(F)) : F \in \mathcal{F}_0\}$$

leading to the smallest asymptotic variance. Values Q_1 and Q_2 are thus for our purposes indistinguishable if $D(O|G_0, Q_1) = D(O|G_0, Q_2)$ with probability one. For $\sqrt{n}(\hat{\mu} - \mu)$ to converge in law to a Gaussian distribution, we generally

need that $D(\cdot|\hat{G}, \hat{Q})$ converges to some element of function class $\mathcal{D}$, and that this is a P_0 -Donsker class. The rate of this convergence will generally not impact the asymptotics for parameter estimate $\hat{\mu}$. When $D(\cdot|\hat{G}, \hat{Q})$ converges to the optimal element in $\mathcal{D}$, two consequences are of note:

1. Our estimator will be locally efficient, so will be asymptotically equivalent to standard locally efficient estimators if the full data working model is correctly specified. But in the more frequent misspecified model scenario, our estimator's asymptotic variance will be equal or superior (modulo improvements due to estimating the coarsening mechanism).
2. In a covariate adjustment problem, we can often choose our working model to ensure a standard unadjusted estimator ignoring covariates is asymptotically equivalent to $\frac{1}{n} \sum_{i=1}^n D(O_i|G_0, Q)$ for some $Q \in \mathcal{Q}$. Our estimator will then guarantee at least the efficiency of the unadjusted technique. This property is not shared by prevailing locally efficient estimators, as we will see in an example.

Our procedure is related to empirical risk minimization methods used for estimating irregular parameters such as regression functions, densities, or Bayes classifiers. If the goal is to select Q in an index set $\mathcal{Q}$ to minimize a risk function $R : \mathcal{Q} \rightarrow \mathbb{R}$, one first represents the risk function as the population mean $E_{P_0}[L(O, Q)]$ of a loss function, ordinarily measuring the error of Q as a predictor of some feature of the observed data O . When the risk $R(Q)$ isn't available, the idea is to use empirical risk $\frac{1}{n} \sum_{i=1}^n L(O_i, Q)$ as a surrogate. We have proposed using the loss function $L(O, Q|G_0) = D^2(O|G_0, Q)$ (depending on the coarsening mechanism G_0) because it is associated with a risk corresponding to asymptotic variance. Hence, selecting $Q \in \mathcal{Q}$ is an intermediate step in applying the estimator $\hat{\mu} = \frac{1}{n} \sum_{i=1}^n D(O_i|\hat{G}, Q)$, and we estimate $\hat{Q} \in \mathcal{Q}$ using empirical risk minimization. Locally efficient estimators maximizing likelihood over working model $\mathcal{F}_0$ implicitly use empirical risk minimization with loss function $L(O, Q(F)|G_0) = -\log dP_{F, G_0}(O)$, which is less targeted toward the parameter of interest. Because our risk function is chosen to optimize the asymptotic efficiency of our resulting parameter estimate, we call our method empirical efficiency maximization. Benefits are best seen through examples, which we now provide.

3 Randomized Experiments with Covariates

We now return to the problem considered in the introduction. Recall the observed data is

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

where W is a baseline covariate vector, $\Delta \in \{0, 1\}$ is a treatment or control indicator, and Y is an observed response. We cast the problem in the counterfactual outcome formulation of Neyman (1923) and Rubin (1974), and suppose a subject's unavailable full data is

$$X = (W, Y_T, Y_C) \sim F_0 \in \mathcal{F},$$

where Y_T and Y_C are the responses that would have occurred under treatment or control. Only one of Y_T or Y_C is ever seen for each subject. Treatment is randomized, with the randomization possibly depending on the covariate measurements in that

$$\pi_0(W) \equiv P(\Delta = 1|W) = P(\Delta = 1|W, Y_T, Y_C). \quad (7)$$

The randomization function $\pi_0(\cdot)$ is known by design, and represents the coarsening mechanism G_0 previously discussed. It is typically assumed that

$$\pi_0(W) \text{ is almost surely bounded away from zero and one,} \quad (8)$$

meaning covariates cannot preclude a subject from being in either study arm. We will make the additional assumption that

$$Y_T \text{ and } Y_C \text{ have finite variance.} \quad (9)$$

For exposition, we assume the data $\{O_i\}_{i=1}^n$ represents an i.i.d. sample, meaning subjects are assigned to treatment or control based on a weighted coin flip, but subjects' coins don't influence one another. Friedman, Furberg, and DeMets (1998) note in clinical trials that such "simple randomization is not often used, even for large studies," because by chance there can be "a serious imbalance in the number of participants assigned to each group." A more realistic scheme would ensure a fixed proportion of subjects were assigned to treatment. Empirical efficiency maximization must be slightly modified for such sampling, and is treated in Rubin and van der Laan (2008).

3.1 Parameters of Interest

We denote the expected responses for subjects assigned to treatment or control with

$$\begin{aligned}\mu_T &= E[Y_T] = E_W[E[Y|\Delta = 1, W]] \\ \mu_C &= E[Y_C] = E_W[E[Y|\Delta = 0, W]].\end{aligned}$$

These two quantities are often, but not always, the parameters of primary importance. One might also express interest in a treatment effect, such as the intention-to-treat parameter

$$\mu_{ITT} = \mu_T - \mu_C.$$

More generally, a treatment effect can be written as $f(\mu_T, \mu_C)$. Particularly in experiments with binary outcomes, interest often lies in relative risks or odds ratios, or their logarithms

$$\begin{aligned}f_{\text{RR}}(\mu_T, \mu_C) &= \mu_T / \mu_C \\ f_{\log(\text{RR})}(\mu_T, \mu_C) &= \log f_{\text{RR}}(\mu_T, \mu_C) \\ f_{\text{OR}}(\mu_T, \mu_C) &= \frac{\mu_T}{1 - \mu_T} / \frac{\mu_C}{1 - \mu_C} \\ f_{\log(\text{OR})}(\mu_T, \mu_C) &= \log f_{\text{OR}}(\mu_T, \mu_C).\end{aligned}$$

3.2 Estimating μ_T or μ_C

Consider estimating the expected response μ_T for subjects assigned to treatment, and note that μ_C can be handled with complete symmetry. A popular approach is the (1952) Horvitz-Thompson estimator

$$\hat{\mu}_T = \frac{1}{n} \sum_{i=1}^n \frac{Y_i \Delta_i}{\pi_0(W_i)},$$

averaging responses of subjects assigned to treatment, but weighing by inverse probabilities of receiving treatment. While simple and unbiased, the estimator can be inefficient, because it ignores covariate information from subjects assigned to control. When $\pi_0(W)$ is constant so that randomization does not depend on baseline measurements, the method leads to an unadjusted estimator making no use of covariate information.

The efficient influence curve for μ_T is given in van der Laan and Robins (2003) as

$$IC_T(O|\pi_0, Q_T^*, \mu_T) = D_T(O|\pi_0, Q_T^*) - \mu_T,$$

for the doubly robust mapping of $\psi(X) = Y_T$ given by

$$D_T(O|\pi_0, Q_T^*) = \frac{Y\Delta}{\pi_0(W)} + (1 - \frac{\Delta}{\pi_0(W)})Q_T^*(W),$$

where

$$Q_T^*(W) = E[Y_T|W] = E[Y|\Delta = 1, W].$$

An asymptotically efficient estimator could be constructed for this problem, by forming a consistent estimate $\hat{Q}_T^*(\cdot)$, and applying

$$\hat{\mu}_T = \frac{1}{n} \sum_{i=1}^n D_T(O_i|\pi_0, \hat{Q}_T^*).$$

This presents no essential challenge when covariate W can only take on a small number of values, because $Q_T^*(w)$ could be fit with the empirical mean of responses in the treatment group's $\{W = w\}$ stratum. When W is instead a large vector of informative covariates, consistently fitting the regression function $Q_T^*(\cdot)$ would be a potentially difficult function approximation task, susceptible to the curse of dimensionality.

A locally efficient approach is to assume a working model $\mathcal{F}_0$ for the conditional distribution $\mathcal{L}(Y|\Delta = 1, W)$. This induces the working index set of functions

$$\mathcal{Q}_T = \{w \rightarrow E_F[Y|\Delta = 1, W = w] : F \in \mathcal{F}_0\}.$$

When fitting the working model, we obtain a fitted function $\hat{Q}_T(\cdot)$. By the discussed robustness property of the efficient influence curve, it follows that

$$E[D_T(O|\pi_0, Q_T)] = \mu_T$$

for any integrable $Q_T(\cdot)$. That is, we can replace the unknown regression function $Q_T^*(\cdot)$ with any function $Q_T(\cdot)$, and still obtain an unbiased estimator. This can be immediately seen, because the first term in $D(O|\pi_0, Q_T)$ has mean μ_T as in the Horvitz-Thompson estimator, and it is a trivial computation to show the second term has mean zero for any integrable $Q_T(\cdot)$. Because of this robustness property, we can estimate the parameter of interest with $\hat{\mu}_T = \frac{1}{n} \sum_{i=1}^n D(O_i|\pi_0, \hat{Q}_T)$. Note that when $\hat{Q}_T$ is fit from the data, $\hat{\mu}_T$ may no longer be exactly unbiased.

An often overlooked issue is how to fit the working model for the conditional distribution $\mathcal{L}(Y|\Delta = 1, W)$. For example, when the response is binary, the most common working model in the local efficiency literature is the logistic regression model, inducing the working index set

$$\mathcal{Q}_T = \{w \rightarrow \frac{1}{1 + \exp(-\beta^T w)} : \beta\},$$

where we suppress the intercept by including it in the covariate vector. The logistic regression is then traditionally fit via maximum likelihood.

When the working model is correctly specified, maximum likelihood will guarantee the fit $\hat{Q}_T(\cdot)$ converges to the true regression function $Q_T^*(\cdot)$, and hence that the parameter estimate $\hat{\mu}_T$ is asymptotically efficient. But when the working model is incorrect, we only know that the maximum likelihood fit converges to the $Q_T(\cdot) \in \mathcal{Q}_T$ minimizing Kullback-Leibler divergence from $Q^*(\cdot)$. While misspecification does not compromise consistency or asymptotic Gaussianity, there is no guarantee that the method asymptotically selects a beneficial element of working index set $\mathcal{Q}_T$ for estimating the parameter of interest μ_T .

The empirical efficiency maximization approach is to instead target the working model element minimizing the asymptotic variance of the resulting parameter estimator. Examine the following result, proven in the appendix.

Theorem 1. *Under assumptions (7), (8), and (9),*

$$\sigma^2(Q_T) \equiv \text{Var}(D_T(O|\pi_0, Q_T)) = K + E[\Delta \frac{1 - \pi_0(W)}{\pi_0^2(W)} |Y - Q_T(W)|^2]$$

for constant

$$K = -\mu_T^2 + E[|\frac{\Delta Y}{\pi_0(W)}|^2] + E[\frac{1 - \pi_0(W)}{\pi_0(W)} \{E[Y_T|W]^2 - (Y - E[Y_T|W])^2\}]$$

not depending on Q_T .

The theorem reveals that the optimal function $Q_T(\cdot)$ minimizes not a Kullback-Leibler divergence, but a weighted mean squared error. We can thus fit a model for $E[Y|\Delta = 1, W]$ with nonlinear least squares and weights proportional to $\Delta_i \frac{1 - \pi_0(W_i)}{\pi_0(W_i)}$. In the R language, this can be accomplished with the `nls()` function. That is, supposing our working model for $\mathcal{L}(Y|\Delta = 1, W)$ is parametrized by β , we propose choosing

$$\hat{\beta} = \text{argmin}_{\beta} \frac{1}{n} \sum_{i=1}^n \Delta_i [\frac{1 - \pi_0(W_i)}{\pi_0(W_i)}]^2 |Y_i - Q_T^{(\beta)}|^2,$$

and estimating the parameter of interest with $\hat{\mu}_T = \frac{1}{n} \sum_{i=1}^n D(O_i|\pi_0, Q_T^{(\hat{\beta})})$.

As proven in Rubin and van der Laan (2008) (albeit under a slightly different sampling scheme), asymptotic Gaussianity can be guaranteed under very general conditions if $Q_T^{(\hat{\beta})}$ converges to a $Q_T^{(\beta)}$. If $\int |Q_T^{(\hat{\beta})}(w) - Q_T^{(\beta)}(w)|^2 dP(w)$

converges to zero in probability and the working index set $\mathcal{Q}$ is a Donsker class, then the estimator $\hat{\mu}$ will be asymptotically equivalent to the unbiased $\frac{1}{n} \sum_{i=1}^n D_T(O_i|\pi_0, Q^{(\beta)})$ we could have applied with the limiting $Q^{(\beta)}$ known. The result will not even depend on the rate of the $Q_T^{(\hat{\beta})} \rightarrow Q_T^{(\beta)}$ convergence.

When interest lies in the expected response μ_C of subjects assigned to the control group, we can likewise fit a working model $\mathcal{F}_0$ for $\mathcal{L}(Y|\Delta = 0, W)$, inducing working index set

$$\mathcal{Q}_C = \{w \rightarrow E_F[Y|\Delta = 0, W = w] : F \in \mathcal{F}_0\}.$$

We can minimize the weighted squared error $\sum_{i=1}^n (1-\Delta_i) \frac{\pi_0(W_i)}{1-\pi_0(W_i)} |Y_i - Q_C(W_i)|^2$ to form a regression fit, and then apply

$$\hat{\mu}_C = \frac{1}{n} \sum_{i=1}^n D_C(O_i|\pi_0, \hat{Q}_C) = \frac{1}{n} \sum_{i=1}^n \left[\frac{(1-\Delta_i)Y_i}{1-\pi_0(W_i)} + \left(1 - \frac{1-\Delta_i}{1-\pi_0(W_i)}\right) \hat{Q}_C(W_i) \right].$$

3.3 Two-phase Designs

Following their introduction in Neyman (1938), two-phase studies have become popular in epidemiology for measuring prevalences of diseases that are difficult or expensive to diagnose, including mental disorders such as depression and schizophrenia. The sampling scheme has also received attention from survey statisticians, but for exposition we restrict to prevalence estimation. As described by Clayton et al. (1998),

A first-phase sample is drawn from the target population and each individual within this sample is then assessed using a cheap and easy-to-use surrogate disease indicator. On the basis of this measurement, the sample is then stratified and a second-phase subsample is drawn. Every member of this second sample receives an accurate diagnostic evaluation to establish their true disease status.

Two-phase studies fall into the experimental framework under consideration in this section. In our notation, W can represent the information collected on all subjects in phase 1, Δ can represent a binary indicator of phase 2 inclusion, and Y can represent whatever “accurate diagnostic evaluation” is made in phase 2. The parameter of interest is then simply μ_T . Clayton et al. (1998) and Alonzo et al. (2003) mention locally efficient techniques utilizing likelihood-based logistic regression models for nuisance parameter fits, which can be improved with empirical efficiency maximization.

3.4 Treatment Effects

After fitting working models for the conditional distributions $\mathcal{L}(Y|\Delta = 1, W)$ and $\mathcal{L}(Y|\Delta = 0, W)$, we have seen how to estimate μ_T and μ_C . For a treatment effect $\mu = f(\mu_T, \mu_C)$, the natural estimate is then the substitution $\hat{\mu} = f(\hat{\mu}_T, \hat{\mu}_C)$. While the previously given techniques maximize efficiencies for $\hat{\mu}_T$ and $\hat{\mu}_C$, they were not designed with a treatment effect in mind. It might often be the case that this effect is of more intrinsic interest than either of the two counterfactual means. For instance, in studies with binary outcomes, the relative risk is often the first number reported. We can in fact fit the two working models to optimize efficiency for the parameter we care most about.

Consider estimating μ_T and μ_C based on functions $Q_T(\cdot)$ and $Q_C(\cdot)$ as before through $\hat{\mu}_T = \frac{1}{n} \sum_{i=1}^n D_T(O_i|\pi_0, Q_T)$ and $\hat{\mu}_C = \frac{1}{n} \sum_{i=1}^n D_C(O_i|\pi_0, Q_C)$, and estimating $f(\mu_T, \mu_C)$ with the substitution $f(\hat{\mu}_T, \hat{\mu}_C)$. Assuming f is differentiable with respect to both arguments and the second derivative matrix is bounded in a neighborhood of (μ_T, μ_C) , the delta method tells us $f(\hat{\mu}_T, \hat{\mu}_C)$ is asymptotically linear with influence curve

$$\begin{aligned} IC_f(O|\pi_0, Q_T, Q_C, \mu_T, \mu_C) &= IC_T(O|\pi_0, Q_T, \mu_T) \frac{\partial}{\partial \mu_T} f(\mu_T, \mu_C) \\ &+ IC_C(O|\pi_0, Q_C, \mu_C) \frac{\partial}{\partial \mu_C} f(\mu_T, \mu_C). \end{aligned}$$

For the relative risk, odds ratio, and their logarithms, partial derivatives are given by

$$\begin{aligned} \left[\frac{\partial f_{RR}}{\partial \mu_T}, \frac{\partial f_{RR}}{\partial \mu_C} \right](\mu_T, \mu_C) &= [\mu_C^{-1}, -\mu_T \mu_C^{-2}] \\ \left[\frac{\partial f_{\log(RR)}}{\partial \mu_T}, \frac{\partial f_{\log(RR)}}{\partial \mu_C} \right](\mu_T, \mu_C) &= [\mu_T^{-1}, -\mu_C^{-1}] \\ \left[\frac{\partial f_{OR}}{\partial \mu_T}, \frac{\partial f_{OR}}{\partial \mu_C} \right](\mu_T, \mu_C) &= \left[\frac{(1 - \mu_C)}{(1 - \mu_T)^2 \mu_C}, \frac{-\mu_T}{(1 - \mu_T) \mu_C^2} \right] \\ \left[\frac{\partial f_{\log(OR)}}{\partial \mu_T}, \frac{\partial f_{\log(OR)}}{\partial \mu_C} \right](\mu_T, \mu_C) &= [\mu_T^{-1} + (1 - \mu_T)^{-1}, -\mu_C^{-1} - (1 - \mu_C)^{-1}]. \end{aligned}$$

Suppose the working models for $\mathcal{L}(Y|\Delta = 1, W)$ and $\mathcal{L}(Y|\Delta = 0, W)$ are together parametrized by some vector β . These working models accordingly induce functions $Q_T^{(\beta)}(W) = E_\beta[Y|\Delta = 1, W]$ and $Q_C^{(\beta)}(W) = E_\beta[Y|\Delta = 0, W]$. Define the function $h^{(\beta)}(O) =$

$$\frac{\partial}{\partial \mu_T} f(\mu_T, \mu_C) \left(1 - \frac{\Delta}{\pi_0(W)}\right) Q_T^{(\beta)}(W) + \frac{\partial}{\partial \mu_C} f(\mu_T, \mu_C) \left(1 - \frac{1 - \Delta}{1 - \pi_0(W)}\right) Q_C^{(\beta)}(W)$$

and the surrogate response

$$\tilde{Y}(O) = -\frac{\partial}{\partial \mu_T} f(\mu_T, \mu_C) \left[\frac{\Delta Y}{\pi_0(W)} - \mu_T \right] - \frac{\partial}{\partial \mu_C} f(\mu_T, \mu_C) \left[\frac{(1 - \Delta)Y}{1 - \pi_0(W)} - \mu_C \right].$$

The influence curve is then $IC_f(O|\pi_0, Q_T, Q_C, \mu_T, \mu_C) = h^{(\beta)}(O) - \tilde{Y}(O)$.

To evaluate this influence curve as a function of β , we need preliminary estimates of μ_T of μ_C . It might often be convenient to use the Horvitz-Thompson results, although any consistent fits will suffice. We then propose aiming to select β for the purpose of minimizing the asymptotic variance of the substitution $f(\hat{\mu}_T, \hat{\mu}_C)$. As this asymptotic variance is the mean of the squared influence curve, it can be approximated empirically to select desirable $Q_T^{(\beta)}(\cdot)$ and $Q_C^{(\beta)}(\cdot)$. We could simply use a nonlinear least squares routine to minimize $\sum_{i=1}^n |\tilde{Y}(O_i) - h^{(\beta)}(O_i)|^2$.

As with estimating μ_T , asymptotic linearity for $f(\hat{\mu}_T, \hat{\mu}_C)$ will require a $Q_T^{(\hat{\beta})} \rightarrow Q_T^{(\beta)}$ and $Q_C^{(\hat{\beta})} \rightarrow Q_C^{(\beta)}$ convergence, but won't depend on the rates at which $Q_T^{(\hat{\beta})}$ and $Q_C^{(\hat{\beta})}$ approach their limits. Convergence will under general conditions guarantee asymptotic Gaussianity for $[\hat{\mu}_T, \hat{\mu}_C]$, and when the delta method can be applied, also for the treatment effect substitution estimate.

3.5 The Intention-to-Treat Parameter

Recall the intention-to-treat parameter is the difference $\mu_{\text{ITT}} = \mu_T - \mu_C$. When forming $\hat{\mu}_T = \frac{1}{n} \sum_{i=1}^n D(O_i|\pi_0, Q_T)$ and $\hat{\mu}_C = \frac{1}{n} \sum_{i=1}^n D_C(O_i|\pi_0, Q_C)$, one can easily find the substitution estimator's influence curve as

$$\begin{aligned} IC_{\text{ITT}}(O|\pi_0, Q_T, Q_C, \mu_T, \mu_C) &= IC_T(O|\pi_0, Q_T, \mu_T) - IC_C(O|\pi_0, Q_C, \mu_C) \\ &= D_T(O|\pi_0, Q_T) - \mu_T - D_C(O|\pi_0, Q_C) + \mu_C \\ &= \left(\frac{\Delta}{\pi_0(W)} - \frac{1 - \Delta}{1 - \pi_0(W)} \right) (Y - Q(W)) - \mu_{\text{ITT}}, \end{aligned}$$

for

$$Q(W) = (1 - \pi_0(W))Q_T(W) + \pi_0(W)Q_C(W).$$

Parametrize working models for $\mathcal{L}(Y|\Delta = 1, W)$ and $\mathcal{L}(Y|\Delta = 0, W)$ by a vector β . The working models then induce index set

$$\mathcal{Q} = \{Q^{(\beta)}(w) = (1 - \pi_0(W))E_\beta[Y|\Delta = 1, w] + \pi_0(W)E_\beta[Y|\Delta = 0, w] : \beta\}.$$

Our representation of the influence curve then tells us the substitution estimator's asymptotic variance is monotone in

$$\sigma^2(\beta) = E\left[\left|\frac{\Delta}{\pi_0(W)} - \frac{1 - \Delta}{1 - \pi_0(W)}\right|^2 |Y - Q^{(\beta)}(W)|^2\right].$$

This is simply a weighted mean squared error for prediction $Q^{(\beta)}(W)$ of response Y , with weight $|\frac{\Delta}{\pi_0(W)} - \frac{1-\Delta}{1-\pi_0(W)}|^2$. That is, the i^{th} subject will be given weight $\pi_0^{-2}(W_i)$ if assigned to the treatment group and $(1 - \pi_0(W_i))^{-2}$ if assigned to the control group. Note that even when randomization does not depend on baseline covariates, so that $\pi_0(W) = \pi_0$, we propose giving different weights to subjects in the two groups in fitting an outcome regression model, and this is not necessarily standard practice. We must solve a weighted nonlinear least squares problem to find

$$\hat{\beta} = \operatorname{argmin}_{\beta} \frac{1}{n} \sum_{i=1}^n \left| \frac{\Delta_i}{\pi_0(W_i)} - \frac{1 - \Delta_i}{1 - \pi_0(W_i)} \right|^2 |Y_i - Q^{(\beta)}(W_i)|^2.$$

Preliminary estimates for μ_T and μ_C are not required.

If the intention-to-treat parameter is the primary quantity of interest, we don't necessarily even have to think in terms of a substitution estimator. We can simply use weighted least squares regression to find some function $\hat{Q}(\cdot)$ that is a good predictor of response Y from covariate W , and estimate the parameter of interest μ_{ITT} with

$$\hat{\mu}_{ITT} = \frac{1}{n} \sum_{i=1}^n \left[\frac{\Delta_i}{\pi_0(W_i)} - \frac{1 - \Delta_i}{1 - \pi_0(W_i)} \right] [Y_i - \hat{Q}(W_i)].$$

For instance, in an experiment with a binary response, we could note the optimal function for asymptotic variance is $Q^*(W) = (1 - \pi_0(W))Q_T^*(W) + \pi_0(W)Q_C^*(W)$, with range between zero and one. Rather than building working models for $\mathcal{L}(Y|\Delta = 1, W)$ and $\mathcal{L}(Y|\Delta = 0, W)$, we could approximate the function directly with a logistic link, inducing the working index set

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

Empirical efficiency maximization for the special case of the intention-to-treat parameter is discussed more generally in Rubin and van der Laan (2008).

3.6 Notes on Convergence

Considering estimation of the expected response μ_T in the treatment group, recall our procedure necessitates solving a weighted nonlinear least squares problem. For simplicity, assume $\pi_0(W) = \pi_0$, so the randomization mechanism doesn't depend on baseline covariate measurements. If using a logistic

regression working model for $\mathcal{L}(Y|\Delta = 1, W)$, we must solve

$$\hat{\beta} = \operatorname{argmin}_{\beta} \frac{1}{n} \sum_{\{i: \Delta_i=1\}} \left| Y_i - \frac{1}{1 + \exp(-\beta^T W_i)} \right|^2. \quad (10)$$

Unfortunately, the problem is not convex in β . Practical software packages will generally have to settle for a local optimum. Working in a neural network context, Sontag and Sussmann (2001) show similar optimization problems can have multiple local minima, raising several issues concerning our estimator's asymptotic behavior.

For $\hat{\mu}_T$ to be asymptotically linear, we generally need that $D_T(\cdot|\pi_0, Q_T^{(\hat{\beta})})$ converges in $L^2(P_0)$ to some function $D_T(\cdot|\pi_0, Q_T)$. This will not be hard to prove if we have a $\hat{\beta} \rightarrow \beta$ convergence in probability or an appropriate $Q_T^{(\hat{\beta})} \rightarrow Q_T$ convergence. However, if $\hat{\beta}$ oscillates between local minima of $E[D_T^2(O|\pi_0, Q_T^{(\beta)})]$, asymptotic Gaussianity could fail, or be hard to establish.

Suppose the process $\beta \rightarrow \frac{1}{n} \sum_{i=1}^n D_T^2(O_i|\pi_0, Q_T^{(\beta)})$ converges in the Glivenko-Cantelli sense to the function $\beta \rightarrow E[D_T^2(O|\pi_0, Q_T^{(\beta)})]$, and let β_0 denote the starting value for the optimization routine used in empirical efficiency maximization. We might hope our $\hat{\beta}$ converges to the β one would obtain from going “downhill” using gradient descent on $\beta \rightarrow E[D_T^2(O|\pi_0, Q_T^{(\beta)})]$, after starting from β_0 , 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, if $\hat{\beta}$ oscillates between different local minima, it is conceivable that the consistency and asymptotic Gaussianity of $\hat{\mu}_T$ could fail.

Work is ongoing to determine what modifications can be made, if any, to ensure a $\hat{\beta}$ convergence when fitting a logistic regression model with empirical efficiency maximization. Until then, our heuristic argument is that standard optimization algorithms seem to force a convergence in all examples we've tried, and counterexamples should only arise in pathological cases.

A $\hat{\beta}$ convergence is not necessary for our class of estimators to be asymptotically well behaved, because $D_T(\cdot|\pi_0, Q_T^{(\hat{\beta})})$ can converge without this condition. For example, two different β 's could lead to equivalent $Q_T^{(\beta)}$'s if there is collinearity. When estimating treatment effects, a $Q_T^{(\hat{\beta})}$ or $Q_C^{(\hat{\beta})}$ convergence is not even always needed. Recall that $IC_{ITT}(O|\pi_0, Q_T, Q_C, \mu_T, \mu_C) = IC_{ITT}(O|\pi_0, Q'_T, Q'_C, \mu_T, \mu_C)$ if $(1 - \pi_0(W))Q_T(W) + \pi_0(W)Q_C(W) = (1 - \pi_0(W))Q'_T(W) + \pi_0(W)Q'_C(W)$. Hence, as long as the correct weighted combination of $Q_T^{(\hat{\beta})}$ and $Q_C^{(\hat{\beta})}$ converges, we can apply $\hat{\mu}_{ITT} = \hat{\mu}_T - \hat{\mu}_C$ and expect appropriate behavior. If encountering software warnings when implementing

our covariate adjustment procedures, one should check whether the estimated influence curve is failing to converge, and not merely that the influence curve parametrization is unidentifiable.

3.7 Numerical Asymptotic Efficiency Calculations

We assessed estimator performance for $\hat{\mu}_T$ in artificial randomized experiments with binary responses, corresponding to the type of two-phase study discussed in Section 3.3. Data structures were generated according to the following mechanism:

$$W \sim N(0, 1)$$

$$\pi_0(W) = P(\Delta = 1|W) = \frac{1}{1 + \exp(-W)} \text{ truncated to be in } [.1, .9]$$

$$P(Y = 1|\Delta = 1, W) = \frac{1}{1 + \exp(-W - \eta W^2)}.$$

Here η determined misspecification of the working logistic regression model for $P(Y = 1|\Delta = 1, W)$, and was varied between 0 and 3 in steps of 0.1. For each value of η , we found the limiting logistic regression coefficients for both maximum likelihood and empirical efficiency maximization estimators based on data generated with a sample of size $n = 100,000$. This sample was also used to find the true μ_T . Based on a new sample of size $n = 100,000$, we then computed asymptotic variances by evaluating $\frac{1}{n} \sum_{i=1}^n D^2(O_i|\pi_0, Q_T) - \mu_T^2$. We considered Q_T corresponding to the Horvitz-Thompson estimator, the locally efficient estimator based on a logistic regression MLE, the empirical efficiency maximizer for this working logistic regression model, and the efficient $Q_T^*(W) = P(Y = 1|\Delta = 1, W)$.

Results are shown in Figure 1, as the misspecification parameter η increased from zero and the working logistic regression model became less appropriate. It was clear that the Horvitz-Thompson estimator was far less efficient than the three estimators making use of covariate information in both study arms. Also clear was that empirical efficiency maximization led to a better logistic regression fit than the MLE for the parameter of interest, as performance closely tracked that of the efficient estimator.

3.8 Other Literature

The classical method for covariate adjustment in randomized experiments is ANCOVA, which reduces to a linear model for predicting response Y from

Figure 1: 1 – 4 represent the Horvitz-Thompson estimator that ignores covariate measurements in the control group, a locally efficient estimator fitting a misspecified logistic regression model’s coefficients with maximum likelihood, our estimator fitting logistic regression coefficients with empirical efficiency maximization, and an asymptotically efficient estimator using empirical mean $\hat{\mu}_T = \frac{1}{n} \sum_{i=1}^n D(O_i | \pi_0, Q_T^*)$ at the true $Q_T^*(W) = E[Y | \Delta = 1, W]$.

covariates W and treatment indicator Δ . With a slightly different sampling scheme than considered here, Freedman (2007) discusses classical methods from an asymptotic efficiency standpoint, and shows adjusted estimators can be inferior to unadjusted ones under model misspecification. For binary outcomes, covariate adjustment traditionally proceeds with logistic modeling. Robinson and Jewell (1991) demonstrate asymptotic efficiency costs for naive adjustment, and Freedman (2008) provides more commentary. Koch et al. discusses covariate adjustment in the context of clinical trials, and Pocock et al. (2002) survey the literature in this setting. Under the local efficiency framework of Robins and Rotnitzky (1992) and subsequent papers, adjustment in experiments is also considered by Tsiatis et al. (2007) and Moore and van der Laan (2007).

4 Covariate Adjustment in Survival Analysis

Let’s move to one of the most common tasks for a biostatistician: estimating a failure time distribution in the presence of right censoring. In typical studies, covariate information is collected on subjects in addition to failure and

censoring time measurements, and our data structure is n i.i.d. copies of

$$O = (W, \Delta = I(T \leq C), \tilde{T} = \min(T, C)).$$

Here $W \in \mathbb{R}^p$ is a baseline covariate vector, T is a failure time, and C is a censoring time. The unavailable full data would be $X = (T, W) \sim F_0$, which we might not directly observe because some subjects cannot be monitored until failure. We will let $G_0(\cdot) = P(C > \cdot)$ denote the survival curve of censoring variable C , and refer to this function as the censoring mechanism or coarsening mechanism.

4.1 Parameters of Interest

Suppose interest lies in estimating the univariate full data parameter,

$$\mu = E[\psi(X)] = E[\psi(T, W)].$$

Pointers for more general parameters will be given in Sections 4.8 and 5. To estimate the survival probability at a single time t , such as when using a five-year survival endpoint, we could take $\psi(T, W) = I(T > t)$. The well-known Kaplan-Meier (1958) estimator ignores covariate information, and would likely be used in practice. While convenient, ignoring informative covariates can lead to a serious loss in efficiency, and anyone who has viewed the estimator's confidence bands can attest that its precision often leaves much to be desired. Clinically predictive measurements provide extra information about failure times of individuals lost to censoring, and thus can enhance estimation. For the mean failure time, we would take $\psi(T, W) = T$, or perhaps the truncated version $\psi(T, W) = T \wedge \tau$ for identifiability. The failure time is also often placed on the log scale.

4.2 Assumptions

The Kaplan-Meier estimate of marginal survival requires the assumption that a subject's censoring process is independent of their failure time, written as

$$T \perp C.$$

We make the additional assumption that censoring is not only independent of failure time, but also of covariates, in that

$$(W, T) \perp C. \tag{11}$$

Our rationale is that if covariates are clinically informative, and censoring is unrelated to the survival outcome, censoring should also be independent of baseline measurements. If a component of covariate vector W is related to censoring but not survival, it can be discarded to satisfy (11). This assumption places the problem within the present work's scope, because it allows the coarsening mechanism to be efficiently estimated with the Kaplan-Meier fit.

When survival at time t is of interest, we can of course truncate failure time T at any point τ larger than t without affecting the parameter of interest. Hence, for identifiability we assume

$$\text{there is a } \tau \text{ such that } \tau \geq T \text{ almost surely and } G_0(\tau) > 0, \quad (12)$$

meaning censoring cannot in principle prevent us from identifying the marginal failure time distribution. Finally, we will assume that

$$(X) = \psi(T, W) \text{ has finite variance.} \quad (13)$$

4.3 Estimation

The efficient influence curve for parameter $\mu = E[\psi(T, W)]$ is

$$IC(O|G_0, Q^*, \mu) = D(O|G_0, Q^*) - \mu,$$

for the doubly robust mapping of $\psi(X) = \psi(T, W)$ given by

$$D(O|G_0, Q^*) = \frac{\Delta\psi(\tilde{T}, W)}{G_0(\tilde{T}_-)} + \int \frac{Q^*(u, W)}{G_0(u)} dM(u|G_0),$$

where

$$Q^*(u, W) = E[\psi(T, W)|T > u, W].$$

Here

$$\begin{aligned} M(u|G_0) &= N(u) - A(u|G_0) \\ N(u) &= I(\tilde{T} \leq u, \Delta = 0) \\ A(u|G_0) &= \int_{-\infty}^u I(\tilde{T} \geq s) \frac{dG_0(s)}{G_0(s_-)}. \end{aligned}$$

The martingale $M(\cdot|G_0)$ is built from the Doob-Meyer decomposition of the counting process $N(\cdot)$ jumping at an observed censoring time, and $A(\cdot|G_0)$ is the right-continuous compensator. To our knowledge, this influence curve representation was derived and interpreted in a series of papers, beginning with Robins and Rotnitzky (1992).

Asymptotically efficient estimation is possible, but would entail consistently estimating the function $Q^*(\cdot, \cdot)$, and along with a consistent estimate of the censoring mechanism $\hat{G}_0(\cdot)$, plugging-in $\hat{\mu} = \frac{1}{n} \sum_{i=1}^n D(O_i | \hat{G}_0, \hat{Q}^*)$. For covariate vector of moderate dimension, consistently estimating $Q^*(\cdot, \cdot)$ would be impractical.

The locally efficient technique is to instead assume a working model $\mathcal{F}_0$ for the conditional failure time distribution $\mathcal{L}(T|W)$, inducing the working index set

$$\mathcal{Q} = \{Q(u, w) = E_F[\psi(T, W) | T > u, W = w] : F \in \mathcal{F}_0\}.$$

After fitting the working model, we are left with a fit $\hat{Q}$. For example, consider Cox's (1972) proportional hazards model, and suppose the parameter of interest is the t -year survival $\mu = P(T > t)$. Parametrized by a coefficient vector β and baseline cumulative hazard $\Lambda_0(\cdot)$, the model induces working index set

$$\mathcal{Q} = \{Q^{(\beta, \Lambda_0)}(u, W) = \exp(-e^{\beta^T W} [\Lambda_0(t \vee u) - \Lambda_0(u)]) : \beta \in \mathbb{R}^p, \Lambda_0(\cdot)\}. \quad (14)$$

The robustness property of the efficient influence curve implies

$$E[D(O|G_0, Q)] = \mu$$

for any function $Q(\cdot, \cdot)$ induced by a working model, whether or not the model is correctly specified. One can see this by noticing that the first term in $D(O|G_0, Q)$ is an unbiased inverse probability of censoring weighted function, as discussed in Robins and Rotnitzky (2005), while the martingale term has mean zero for any $Q(\cdot, \cdot)$ induced by a $\mathcal{L}(T|W)$ working model.

After posing such a working model, we are left with the task of fitting. The Cox model is ordinarily fit with Cox's partial likelihood technique, but it is not clear that a likelihood-based approach would be optimal for the parameter of interest. Rather than targeting the working model element minimizing Kullback-Leibler divergence from the true unknown data generating distribution, we can aim to optimize asymptotic variance by minimizing empirical risk $\hat{R}(Q) = \frac{1}{n} \sum_{i=1}^n D^2(O_i | \hat{G}_0, Q)$ over working index set $\mathcal{Q}$ to form $\hat{Q}$, and then apply parameter estimate $\hat{\mu} = \frac{1}{n} \sum_{i=1}^n D(O_i | \hat{G}_0, \hat{Q})$.

As in the previous section, it can be shown under extremely general conditions that an appropriate $\hat{Q} \rightarrow Q$ convergence implies $\sqrt{n}(\hat{\mu} - \mu)$ converges in law to a Gaussian distribution, regardless of the $\hat{Q} \rightarrow Q$ rate.

4.4 Weighted Squared Error Loss Function

As in the randomized experiment example, we can conveniently represent the empirical efficiency maximization approach as a model fitting method utilizing

a weighted least squares criterion. To ease exposition, in this section we'll assume that

$$\text{censoring time } C \text{ has a continuous distribution.} \quad (15)$$

The following theorem represents the variance of $\hat{\mu} = \frac{1}{n} \sum_{i=1}^n D(O_i|G_0, Q)$, which we could apply as an estimator if the censoring mechanism $G_0(\cdot)$ were known.

Theorem 2. *Under assumptions (11), (12), (13), and (15), we have that*

$$\begin{aligned} \sigma^2(Q) &= \text{Var}(D(O|G_0, Q)) = E[D^2(O|G_0, Q)] - \mu^2 \\ &= K + E_{(O,U) \sim \mathcal{L}(O) \times G_0} \left[\frac{I(\tilde{T} \geq U)}{G_0^3(U)} \left| \frac{\Delta\psi(\tilde{T}, W)G_0(U)}{G_0(\tilde{T})} - Q(U, W) \right|^2 \right], \end{aligned}$$

for constant

$$K = -\mu^2 + E \left[\frac{\Delta\psi^2(\tilde{T}, W)}{G_0^2(\tilde{T})} + E_{(O,U) \sim \mathcal{L}(O) \times G_0} \left[\frac{I(\tilde{T} \geq U)}{G_0^3(U)} \left| \frac{\Delta\psi(\tilde{T}, W)G_0(U)}{G_0(\tilde{T})} \right|^2 \right] \right]$$

not depending on $Q(\cdot, \cdot)$.

The result indicates that the working survival model should not necessarily be fit to maximize a likelihood or partial likelihood, but to minimize a special weighted mean squared error. The expectation of interest can be approximated empirically, because we can approximate $\mathcal{L}(O)$ with the empirical distribution on $\{O_i\}_{i=1}^n = \{W_i, \Delta_i, \tilde{T}_i\}_{i=1}^n$, and approximate the censoring mechanism $G_0(\cdot)$ with our fit $\hat{G}_0(\cdot)$. Evaluating empirical risk $\hat{R}(Q) = \frac{1}{n} \sum_{i=1}^n D^2(O_i|\hat{G}_0, Q)$ then necessitates computing a double integral, which can be accomplished through Monte Carlo. We could then attempt to minimize this empirical risk over the working index set $\mathcal{Q}$ induced by the working survival model, and aim to be more asymptotically efficient than someone fitting such a model through standard techniques. The algorithm can be stated as follows:

1. Draw $\{w_b, \delta_b, \tilde{t}_b\}_{b=1}^B$ from $\{W_i, \Delta_i, \tilde{T}_i\}_{i=1}^n$ with replacement.
2. Draw B i.i.d. replicates $\{u_b\}_{b=1}^B$ from censoring mechanism fit $\hat{G}_0$.
3. Form weights $a_b = I(\tilde{t}_b \geq u_b) / \hat{G}_0^3(u_b)$ for $b = 1, \dots, B$.
4. Form surrogate responses $y_b = \delta_b \psi(\tilde{t}_b, w_b) \hat{G}_0(u_b) / \hat{G}_0(\tilde{t}_b)$ for $b = 1, \dots, B$.
5. Choose $Q \in \mathcal{Q}$ to minimize the weighted squared error $\frac{1}{B} \sum_{b=1}^B a_b |y_b - Q(u_b, W_b)|^2$.

For working index set $\mathcal{Q}$ induced by a Cox model, empirical efficiency maximization may appear to require solving an infinite-dimensional optimization problem, because we must minimize our criterion over all baseline hazard functions. However, inspection of the algorithm reveals that $Q(u, W)$ only needs to be evaluated at the censoring times $\{c_b\}_{b=1}^B$, so the problem becomes finite-dimensional, albeit nonlinear and unorthodox.

4.5 Gain from Estimating the Censoring Mechanism

Given a function Q and censoring mechanism fit $\hat{G}_0(\cdot)$, it would be useful to have a representation for the influence curve of $\hat{\mu} = \frac{1}{n} \sum_{i=1}^n D(O_i | \hat{G}_0, Q)$. Under independence assumption (11), we can efficiently estimate the censoring mechanism according to Kaplan-Meier, and as previously discussed will only increase our asymptotic precision relative to $\hat{\mu} = \frac{1}{n} \sum_{i=1}^n D(O_i | G_0, Q)$.

It can be shown through Theorem 2.3 of van der Laan and Robins (2003) that the influence curve is of the form $IC(O | G_0, \tilde{Q}, \mu) = D(O | G_0, \tilde{Q}) - \mu$, for

$$\tilde{Q}(u, W) = Q(u, W) + \phi(u).$$

Here $\phi(\cdot)$ is the function added to $Q(\cdot, \cdot)$ that minimizes asymptotic variance.

Given $Q(\cdot, \cdot)$, the function $\phi(\cdot)$ can be approximated from the data through a linear regression. To see this, note that the first argument of $Q(\cdot, \cdot) + \phi(\cdot)$ only has to be evaluated at the unique censoring times in $D(O | \hat{G}_0, Q + \phi)$. To this end, let ϕ_j denote the value of $\phi(\cdot)$ at the j^{th} unique censoring time.

When using the discrete Kaplan-Meier estimate for the censoring mechanism, the integral $\int \frac{Q(c, W) + \phi(c)}{\hat{G}_0(c)} dM(c | \hat{G}_0)$ becomes a finite sum with at most p terms, where p is the number of unique censoring times. Examining our representation of $D(O | \hat{G}_0, Q + \phi)$, we can therefore clearly form surrogate responses y_i and covariate vectors $x_i = [x_{i,1}, \dots, x_{i,p}]$ to write

$$\hat{R}(Q + \phi) = \frac{1}{n} \sum_{i=1}^n D^2(O_i | \hat{G}_0, Q + \phi) = \frac{1}{n} \sum_{i=1}^n \left| y_i - \sum_{j=1}^p x_{i,j} \phi_j \right|^2.$$

Among all functions $\phi(\cdot)$, we can thus approximate the asymptotic variance minimizer by solving for $[\hat{\phi}_1, \dots, \hat{\phi}_p]$ using linear least squares.

Consequently, we can empirically approximate the asymptotic variance of estimator $\hat{\mu} = \frac{1}{n} \sum_{i=1}^n D(O_i | \hat{G}_0, Q)$ through $\hat{R}(Q + \hat{\phi}) = \frac{1}{n} \sum_{i=1}^n D^2(O_i | \hat{G}_0, Q + \hat{\phi})$, whose evaluation requires a linear regression. For extra notation, let $\hat{R}(Q)$ denote this empirical asymptotic variance estimate taking the fitted censoring mechanism into account.

4.6 Low-dimensional Fluctuation Method

A practical covariate adjustment approach is to begin with some Q , such as that corresponding to the unadjusted estimate ignoring covariate information, and consider a set $\mathcal{Q}$ of low-dimensional fluctuations making use of the covariates. We could then attempt to find the optimal element of working index set $\mathcal{Q}$ through a grid search, and avoid a more difficult nonlinear minimization over a large working survival model.

Suppose interest lies in t -year survival $\mu = P(T > t)$. Let $\Lambda^{(0)}(\cdot)$ denote the marginal cumulative hazard function for survival time. Also, let $V(W)$ denote some reduction of covariate vector W , such as a linear combination found through principal components analysis or a Cox model. We could then form the fluctuations,

$$\Lambda^{(\beta)}(u|W) = \Lambda^{(0)}(u)e^{\beta^T V(W)}.$$

This would induce working index set,

$$\mathcal{Q} = \{Q^{(\beta)}(u, W) = \exp(-e^{\beta^T V(W)}[\Lambda^{(0)}(t \vee u) - \Lambda^{(0)}(u)]) : \beta\},$$

parametrized only by the low-dimensional β . To work with members of $\mathcal{Q}$, note that the marginal cumulative hazard can be approximated with the Nelson-Aalen estimate

$$\hat{\Lambda}^{(0)}(u) = \frac{1}{n} \sum_{i=1}^n \int_{-\infty}^u dI(\Delta_i = 0, \tilde{T}_i \leq s).$$

Following the previous subsection, we could then approximate asymptotic variances resulting from elements of $\mathcal{Q}$ through the empirical $\hat{R}(Q^{(\beta)})$. For low-dimensional β , this can be minimized via a grid search to form $\hat{\beta}$. Subject to the same convergence issues arising in Section 3.6, we would then hope $Q^{(\hat{\beta})}$ converges to the optimal element within our working index set.

Because it can be shown that $\beta = 0$ corresponds to the Q appearing the Kaplan-Meier influence curve, this use of covariate adjustment can only help us asymptotically. As we will soon see, the same cannot be said for locally efficient methods fitting working models through standard likelihood-based techniques.

4.7 Numerical Asymptotic Efficiency Calculations

We explored estimation techniques by generating data structures as follows:

$$\begin{aligned} W &\sim \text{Uniform}(0, 1) \\ \{T|W\} &\sim N\left(\frac{10}{1 + \exp(-\eta(W - \frac{1}{2}))}, 2.5^2\right) \\ P(C = 3) &= P(C = \infty) = \frac{1}{2}. \end{aligned}$$

The value $\eta = 0$ corresponded to a null model in which the covariate W was completely uninformative for survival, which was a special case of the Cox model. Hence, η could be considered a $\mathcal{L}(T|W)$ working model misspecification parameter. We varied η from 0 to 3 in steps of 0.1. The censoring variable C corresponded to flipping a fair coin, and either censoring a subject at time 3 or not censoring at all. In this example, we considered the censoring mechanism $G_0(\cdot)$ to be known to the data analyst, who would attempt to estimate the five-year survival rate $\mu = P(T > 5)$.

Evaluating the doubly robust mapping $D(O|G_0, Q)$ only required evaluating $Q(u, W)$ at $u = 3$, so we simplify notation by writing $Q(W)$. One can check that,

$$D(O|G_0, Q) = 2I(\tilde{T} > 5) + [2(1 - \Delta) - I(\tilde{T} \geq 3)]Q(W),$$

and that asymptotic variance $\sigma^2(Q) = E[D^2(O|G_0, Q)] - \mu^2$ is monotone in the weighted squared error $E[I(\tilde{T} \geq 3)|2I(\tilde{T} > 5) - Q(W)|^2]$. Empirical efficiency maximization thus reduced to selecting a working set $\mathcal{Q}$, and finding the $\hat{Q} \in \mathcal{Q}$ to minimize a weighted least squares. Using a Cox model as the working model for the conditional failure time distribution $\mathcal{L}(T|W)$, the working index set $\mathcal{Q}$ was parametrized through

$$\begin{aligned} Q^{(\beta_1, \Lambda_0)}(W) &= P_{\beta_1, \Lambda_0}(T > 5|T > 3, W) = \exp(-e^{\beta_1 W}(\Lambda_0(5) - \Lambda_0(3))) \\ &= \exp(-e^{\beta_0 + \beta_1 W}), \end{aligned}$$

for $\beta_0 = \log(\Lambda_0(5) - \Lambda_0(3))$.

We considered estimators $\hat{\mu} = \frac{1}{n} \sum_{i=1}^n D(O_i|G_0, Q^{(\hat{\beta}_0, \hat{\beta}_1)})$. For each model misspecification parameter η , we evaluated the limiting $(\hat{\beta}_0, \hat{\beta}_1)$ of standard partial likelihood maximization and empirical efficiency maximization, through generating a dataset with sample size $n = 100,000$. The `nls()` function in the R language was used to solve the weighted nonlinear least squares problem in empirical efficiency maximization. This simulated dataset also allowed

Figure 2: 1–4 represent the Kaplan-Meier estimator that ignores covariates, a locally efficient estimator based on a misspecified Cox model fit with the usual partial likelihood technique and Breslow (1974) baseline cumulative hazard, an empirical efficiency maximization estimator based on the Cox model, and the efficient estimator.

us to find survival probability $\mu = P(T > 5)$ via Monte Carlo. Using a new independent sample of the same size, we then computed asymptotic variances for the two estimators by evaluating $\frac{1}{n} \sum_{i=1}^n D^2(O_i | G_0, Q) - \mu^2$. Additionally, we computed asymptotic variances for the Kaplan-Meier influence curve's $Q^{\text{KM}}(W) = P(T > 5 | T > 3)$, and the efficient influence curve's $Q^*(W) = P(T > 5 | T > 3, W)$. The former was a constant, found in the initial Monte Carlo simulation, while the latter was known by design.

Asymptotic variance results for the four estimators are displayed in Figure 2. Surprisingly, we lost precision when attempting to utilize informative covariates and fit a locally efficient Cox model with partial likelihood. Such a technique appeared worse than ignoring covariates altogether, and using the Kaplan-Meier estimator. When the working Cox model was instead fit with empirical efficiency maximization, performance greatly improved, and we saw that covariate W enhanced estimation. An interesting phenomenon is that even with a misspecified failure time model, elements of the working model can lead to estimators that are essentially fully efficient for our parameter, and likelihood-based estimates do not always converge to such elements.

4.8 Other Survival Parameters

We have limited discussion to full data population means $\mu = \psi(T, W)$, but other parameters could be of interest. To estimate the cumulative hazard at a point t , one could apply a substitution estimator based on an estimate of marginal survival, and fit a working model to enhance efficiency as in Section 5.1. Estimation of failure time quantiles could proceed with the estimating equation method of Section 5.3. Extensions to multivariate parameters are discussed in Section 5.4, but it is not clear how one would best use a working failure time model to best fit an entire survival curve.

4.9 Comparing Survival Distributions

We now sketch how empirical efficiency maximization could be used for the two-sample problem with right censoring, eschewing formality. Suppose binary $A \in \{0, 1\}$ defines two strata, and can be pulled out of covariate vector W so the observed data is

$$O = (W, A, \Delta = I(T \leq C), \tilde{T} = \min(T, C)).$$

One might like to know whether A is associated with survival, meaning whether the conditional law $\mathcal{L}(T|A = 0)$ differs from $\mathcal{L}(T|A = 1)$.

Let $\Lambda_0(\cdot)$ and $\Lambda_1(\cdot)$ denote the cumulative hazard functions in the two populations. Fleming and Harrington (1991) review how many popular weighted logrank test statistics are $n^{-1/2}$ -scaled versions of

$$\hat{\mu} = \int_a^b \hat{K}(u) d\{\hat{\Lambda}_1(u) - \hat{\Lambda}_0(u)\}, \quad (16)$$

where $\hat{\Lambda}_1(\cdot)$ and $\hat{\Lambda}_0(\cdot)$ are Nelson-Aalen estimates of the cumulative hazards. Testing can be viewed through the lens of parameter estimation, for parameter

$$\mu = \int_a^b K(u) d\{\Lambda_1(u) - \Lambda_0(u)\},$$

where $K(\cdot)$ is a weight function. For the widely used logrank statistic, this weight function is defined through

$$K(u) = P(A = 0)P(A = 1) \frac{P(\tilde{T} > u|A = 1)P(\tilde{T} > u|A = 0)}{P(\tilde{T} > u)},$$

and estimated empirically with $\hat{K}(\cdot)$ in test statistic $\sqrt{n}\hat{\mu}$. Because weighted logrank parameters quantify differences in integrated cumulative hazards, testing $H_0 : \mu = 0$ against $H_A : \mu \neq 0$ specifies a test for equality of survival distributions.

Covariate adjustment involves using baseline measurements $\{W_i\}_{i=1}^n$ to more efficiently estimate parameter μ , and thus increase power. Unfortunately, several complications arise. For one, the parameter of interest is not merely a full data population mean $\mu = E[\psi(T, W)]$. Moreover, we are no longer in this work's primary setting, because μ is not necessarily even a full data parameter, as it can depend on the coarsening mechanism $\mathcal{L}(C)$ in weight function $K(\cdot)$. However, these issues can be finessed.

From fits for conditional laws $\mathcal{L}(T|A = 0, W)$ and $\mathcal{L}(T|A = 1, W)$, the techniques of Section 4.3 could be used in the $\{A = 0\}$ and $\{A = 1\}$ samples to form covariate-adjusted estimates of the marginal survival curves $\hat{S}_1(\cdot)$ and $\hat{S}_0(\cdot)$. These curves could then be mapped into cumulative hazard estimates, through

$$\hat{\Lambda}_j(t) = \int_{-\infty}^t \frac{d\hat{S}_j(u)}{\hat{S}_j(u_-)}.$$

Such cumulative hazard fits could then be plugged into (16) to form a covariate-adjusted estimate of the weighted logrank parameter of interest. If $\sqrt{n}(\hat{\mu} - \mu) \rightarrow N(0, \sigma^2)$ in law, and an estimator of asymptotic variance σ^2 is available from a bootstrap or influence curve calculation, a natural test involves comparing statistic $\sqrt{n}\hat{\mu}/\hat{\sigma}$ to a standard Gaussian distribution. The two-sample problem consequently involves fitting working models for the conditional $\mathcal{L}(T|W, A = 0)$ and $\mathcal{L}(T|W, A = 1)$ distributions in the two populations.

Against fixed alternatives $\mu \neq 0$, virtually any weighted logrank procedure will have power tending to one as the sample size grows. Hence, procedures are traditionally examined in terms of power against local alternatives. The influence curve can still be defined as the function such that $\hat{\mu} - \mu = \frac{1}{n} \sum_{i=1}^n IC(O_i|P) + o_P(n^{-1/2})$, and asymptotic power against local alternatives can be assessed through the variance of the limiting law of $\sqrt{n}(\hat{\mu} - \mu)$. To be completely precise, the parameter μ and data generating distribution P would have to be indexed by sample size n .

We consider the cumulative hazards Λ_0 and Λ_1 approaching with sample size, to be formalized as follows. Noting that the empirical $\hat{K}(\cdot)$ should be $\sqrt{n}$ -consistent for weight function $K(\cdot)$, the two cumulative hazard functions should be perturbed with sample size so that,

$$\int_a^b (\hat{K}(u) - K(u)) d\{\Lambda_1(u) - \Lambda_0(u)\} = o_P(n^{-1/2}).$$

Additionally, the $n^{-1/2}$ convergence of $\hat{K}(\cdot)$ should imply that

$$\int_a^b (\hat{K}(u) - K(u)) d\{(\hat{\Lambda}_1(u) - \hat{\Lambda}_0(u)) - (\Lambda_1(u) - \Lambda_0(u))\} = o_P(n^{-1/2}).$$

Telescoping then yields,

$$\begin{aligned} \hat{\mu} - \mu &= \int_a^b \hat{K}(u) d\{\hat{\Lambda}_1(u) - \hat{\Lambda}_0(u)\} - \int_a^b K(u) d\{\Lambda_1(u) - \Lambda_0(u)\} \\ &= \int_a^b (\hat{K}(u) - K(u)) d\{\Lambda_1(u) - \Lambda_0(u)\} \\ &+ \int_a^b (\hat{K}(u) - K(u)) d\{(\hat{\Lambda}_1(u) - \hat{\Lambda}_0(u)) - (\Lambda_1(u) - \Lambda_0(u))\} \\ &+ \int_a^b K(u) d\{(\hat{\Lambda}_1(u) - \hat{\Lambda}_0(u)) - (\Lambda_1(u) - \Lambda_0(u))\} \\ &= \int_a^b K(u) d\{(\hat{\Lambda}_1(u) - \hat{\Lambda}_0(u)) - (\Lambda_1(u) - \Lambda_0(u))\} + o_P(n^{-1/2}). \end{aligned}$$

Therefore, the influence curve of $\hat{\mu}$ is that of $\int_a^b K(u) d\{\hat{\Lambda}_1(u) - \hat{\Lambda}_0(u)\}$, the estimator we could have applied had we known the weight function $K(\cdot)$. Further, Fubini's theorem tells us

$$\begin{aligned} \int_a^b K(u) d\{\hat{\Lambda}_1(u) - \hat{\Lambda}_0(u)\} &= \int_a^b \left[\int_a^u dK(s) + K(a) \right] d\{\hat{\Lambda}_1(u) - \hat{\Lambda}_0(u)\} \\ &= \int_a^b \left[\int_u^b d\{\hat{\Lambda}_1(s) - \hat{\Lambda}_0(s)\} \right] dK(u) + K(a) \int_a^b d\{\hat{\Lambda}_1(u) - \hat{\Lambda}_0(u)\} \\ &= \{\hat{\Lambda}_1(b) - \hat{\Lambda}_0(b)\}(K(b) - K(a)) - \int_a^b \{\hat{\Lambda}_1(u) - \hat{\Lambda}_0(u)\} dK(u) \\ &+ K(a)\{(\hat{\Lambda}_1(b) - \hat{\Lambda}_0(b)) - (\hat{\Lambda}_1(a) - \hat{\Lambda}_0(a))\} \\ &= K(b)(\hat{\Lambda}_1(b) - \hat{\Lambda}_0(b)) - K(a)(\hat{\Lambda}_1(a) - \hat{\Lambda}_0(a)) - \int_a^b \{\hat{\Lambda}_1 - \hat{\Lambda}_0\} dK. \end{aligned}$$

Let $IC_1^{(t)}$ and $IC_0^{(t)}$ denote the influence curves of the covariate-adjusted marginal survival estimates $\hat{S}_0(t)$ and $\hat{S}_1(t)$, as previously defined. Assuming such estimators are regular and the alternatives are contiguous, we can write $\hat{S}_j(t) - S_j(t) = \frac{1}{n} \sum_{i=1}^n IC_t^{(j)}(O_i) + o_P(n^{-1/2})$, for $j = 0, 1$. The delta method can then be used to show cumulative hazard estimate $\hat{\Lambda}_j(t)$ has influence curve $\frac{IC_j^{(t)}(O)}{S_j(t-)}$. Our representation then gives the influence curve for

logrank parameter estimate $\hat{\mu}$ as

$$\begin{aligned}
 IC(O) = & K(b) \left(\frac{IC_1^{(b)}(O)}{S_1(b_-)} - \frac{IC_0^{(b)}(O)}{S_0(b_-)} \right) - K(a) \left(\frac{IC_1^{(a)}(O)}{S_1(a_-)} - \frac{IC_0^{(a)}(O)}{S_0(a_-)} \right) \\
 & - \int_a^b \left(\frac{IC_1^{(u)}(O)}{S_1(u_-)} - \frac{IC_0^{(u)}(O)}{S_0(u_-)} \right) dK(u).
 \end{aligned}$$

With this influence curve, we could use empirical asymptotic variance estimate $\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n IC^2(O_i)$ when fitting working models for the $\mathcal{L}(T|W, A = 1)$ and $\mathcal{L}(T|W, A = 0)$ distributions. It might be computationally convenient to begin with unadjusted Kaplan-Meier fits for the two conditional laws, which yields the unadjusted weighted logrank test statistic, and consider low-dimensional covariate-adjusted fluctuations as in Section 5.2.

In this section, we've made no assumptions about covariate $A \in \{0, 1\}$ being randomized, and have considered separately fitting cumulative hazards in the $\{A = 1\}$ and $\{A = 0\}$ groups. Under randomization, additional efficiency gains appear possible for estimation of parameter μ . Without going into details, the efficient influence curves for the $S_0(\cdot)$ and $S_1(\cdot)$ survival curves would again involve the conditional distributions $\mathcal{L}(T|A = 0, W)$ and $\mathcal{L}(T|A = 1, W)$. Fits of these conditional distributions would again lead to estimates of $S_0(\cdot)$ and $S_1(\cdot)$, which could be plugged into cumulative hazard estimates, and in turn an estimate of the parameter. The squared influence curve of parameter μ could then be used as a loss function for fitting the $\mathcal{L}(T|A = 0, W)$ and $\mathcal{L}(T|A = 1, W)$ working models.

4.10 Other Literature

Locally efficient estimation in survival analysis was pioneered in Robins and Rotnitzky (1992). Chapter 3 of van der Laan and Robins (2003) provides an overview. Along such lines, Zeng (2004) considers locally efficient estimation of marginal survival based on a combination of Cox modeling and smoothing. Similar methods are given in Scharfstein and Robins (2001) and Satten et al. (2002). It is fair to say that most work in this area has been directed toward using covariates to correct for dependent censoring $\neg\{T \perp C\}$ plaguing Kaplan-Meier, rather than towards increasing efficiency.

Covariate-adjusted methods for comparing survival distributions are treated in the original article of Robins and Rotnitzky. Lu (2006) gave a talk presenting covariate-adjusted tests asymptotically outperforming the logrank procedure, and a forthcoming article of Lu and Tsiatis will apparently expound.

Work in this area has traditionally examined power at alternatives where the $\mathcal{L}(T|A = 0)$ and $\mathcal{L}(T|A = 1)$ distributions obey a proportional hazards assumption, while our sketched approach instead bases tests on estimation of a parameter defined nonparametrically.

5 Extensions

We previously restricted attention to estimating univariate full data population means of the form $\mu = E[\psi(X)] \in \mathbb{R}$. This was meant to be expository, and empirical efficiency maximization in this limited setting can go a long way. Modifications are required to attack more general parameters. In this section we informally sketch how one might proceed, and review several extensions already made.

5.1 Substitution

With treatment effects in randomized experiments, we saw how our approach could handle parameters of the form $f(\mu_1, \dots, \mu_k)$, where μ_j was a full data population mean. The efficient influence curve for μ_j was

$$IC_j(O|G_0, Q_j^*, \mu_j) = D_j(O|G_0, Q_j^*) - \mu_j.$$

We considered estimators $\hat{\mu}_j = \frac{1}{n} \sum_{i=1}^n D_j(O_i|G_0, \hat{Q}_j)$, and applied the substitution $\hat{\mu} = f(\hat{\mu}_1, \dots, \hat{\mu}_k)$. We used the delta method to find the influence curve of such estimators as a function of $(Q_1, \dots, Q_k)$. When a full data working model induced a working index set of such $(Q_1, \dots, Q_k)$, we considered finding the optimal element for our substitution estimator by minimizing the squared influence curve's empirical mean.

5.2 Low-dimensional Working Index Set $\mathcal{Q}$

The computational method of Section 5.2 can be generalized. Recall that we considered an influence curve of the form $IC(O|G_0, Q, \mu) = D(O|G_0, Q) - \mu$, and wanted to select from a working index set $\mathcal{Q}$ to enhance precision. Ordinarily the empirical efficiency maximization idea would be to minimize $\frac{1}{n} \sum_{i=1}^n D^2(O_i|\hat{G}_0, Q)$ over the working index set $\mathcal{Q}$. Unfortunately, this isn't always feasible. Sometimes it is convenient to begin with an initial $Q^{(0)}$, apply some form of dimensionality reduction to the covariates, and consider a low dimensional fluctuation $\mathcal{Q} = \{Q^{(\epsilon)} : \epsilon \in \mathbb{R}\}$. It can then become fairly simple to minimize $\frac{1}{n} \sum_{i=1}^n D^2(O_i|\hat{G}_0, Q^{(\epsilon)})$ via grid search. The procedure

is particularly useful in covariate adjustment problems when taking the initial $Q^{(0)}$ as the element corresponding to a standard unadjusted estimator.

5.3 Estimating Equations

Not all parameters of interest are full data population means, or simple functions of such means. Empirical efficiency maximization can be generalized for the types of parameters considered in van der Laan and Robins (2003): those solving estimating equations. Recall that the full data is defined as $\{X_i\}_{i=1}^n \sim F_0$, while the observed data is the coarsened $\{O_i\}_{i=1}^n \sim P_{F_0, G_0}$. Suppose $\mu = \mu(F_0) \in \mathbb{R}$ is defined as the solution to $0 = E_{F_0}[\psi(X|\rho(F_0), \mu)]$. Here $\rho = \rho(F_0)$ is a nuisance parameter. If the full data were available, and we could approximate the nuisance parameter at a fast enough rate with $\hat{\rho}$, we could estimate μ with the solution $\hat{\mu}$ of $0 = \frac{1}{n} \sum_{i=1}^n \psi(X_i|\hat{\rho}, \mu)$. Under identifiability and regularity conditions, the estimator would be asymptotically linear. In a nonparametric (saturated) full data model, any other asymptotically linear estimator would be asymptotically equivalent to it, so ψ would in a sense be the unique estimating function.

van der Laan and Robins (2003) discuss mapping a full data estimating function ψ into estimating functions suitable for the observed data $\{O_i\}_{i=1}^n$. Suppose $F_0 \in \mathcal{F}$, for $\mathcal{F}$ the full data model. Applying the doubly robust mapping in van der Laan and Robins's Theorem 2.1 gives rise to the estimating function $D(O|G_0, Q(F), \rho(F_0), \mu(F_0))$, with the robustness property that

$$E_{P_{F_0, G_0}}[D(O|G_0, Q(F), \rho(F_0), \mu(F_0))] = 0 \text{ for any } F \in \mathcal{F}.$$

With the observed data, we could then form a parameter estimate $\hat{\mu}$ by solving

$$0 = \frac{1}{n} \sum_{i=1}^n D(O_i|G_0, Q(F), \hat{\rho}, \mu).$$

Under additional identifiability and regularity conditions, van der Laan and Robins note the estimator will be asymptotically linear with influence curve

$$\hat{\mu} = \mu + \frac{1}{n} \sum_{i=1}^n KD(O_i|G_0, Q(F), \rho(F_0), \mu(F_0)) + o_{P_0}(n^{-1/2}),$$

for constant K not depending on the data. As with the simpler parameters already considered, efficiently estimating the coarsening mechanism with $\hat{G}$ from a correctly specified submodel can only help asymptotic variance. The $Q(F)$ minimizing asymptotic variance is $Q(F_0)$, and efficient estimators

can often be constructed by approximating $Q(F_0)$ in the estimating equation. When the function approximation problem becomes too difficult, locally efficient estimation proceeds by fitting F in a working model $\mathcal{F}_0$ for the full data generating distribution, or equivalently Q in the induced working index set $\mathcal{Q} = \{Q(F) : F \in \mathcal{F}_0\}$. A $\hat{Q} \rightarrow Q$ convergence will generally suffice for asymptotic linearity of the parameter estimate when the coarsening mechanism is known or can be correctly modeled, regardless of the $\hat{Q} \rightarrow Q$ rate. We propose empirically targeting such a fit to optimize the precision of $\hat{\mu}$.

It is well known (e.g. Bickel et al. (1998), van der Laan and Robins (2003), Tsiatis (2006)) that the influence curve's constant K is given by

$$K = -\left\{\frac{d}{d\mu}E_{P_0}[D(O|G_0, Q(F), \rho(F_0), \mu)]|_{\mu=\mu(F_0)}\right\}^{-1},$$

and it can be shown that this constant generally does not depend on $Q(F)$. We will not provide a formal proof, but the reasoning is very simple, for those acquainted with efficiency theory in censored data structures. The expectation of $D(O|G_0, Q(F), \rho(F_0), \mu)$ can often be represented as the expectation of an inverse probability of censoring weighted term added to the expectation of a term the augmentation space. The former doesn't depend on $Q(F)$, while the latter is zero for any $\mu \in \mathbb{R}$ and $Q(F)$.

As the constant doesn't depend on $Q(F)$, the asymptotic variance of the estimating equation estimator is monotone in $E[D^2(O_i|G_0, \rho(F_0), \mu(F_0))]$. With a coarsening mechanism fit $\hat{G}$, a preliminary parameter estimate $\hat{\mu}$, and a nuisance parameter estimate $\hat{\rho}$, we could approximate this asymptotic variance with $\frac{1}{n} \sum_{i=1}^n D^2(O_i|\hat{G}, Q(F), \hat{\rho}, \hat{\mu})$, and minimize over a working index set $\mathcal{Q}$.

We should note that not all parameters of interest can be dealt with according to estimating function methodology, even among smooth regular parameters estimable at the parametric $n^{-1/2}$ rate. In the surveys of van der Laan and Robins (2003) and Tsiatis (2006), it is assumed the scaled full data efficient influence curve $\psi(X|\rho(F_0), \mu(F_0))$ can be parametrized by the $\mu(F_0)$ of interest and nuisance parameter $\rho(F_0)$. The $\mu(F)$ and $\rho(F)$ are assumed to be *variation independent* at F_0 , essentially meaning that one can be fixed while the other fluctuates as full data distribution F is perturbed about F_0 . The assumption allows one to actually solve the equation $0 = \frac{1}{n} \sum_{i=1}^n \psi(X_i|\hat{\mu}, \hat{\rho})$ and form a parameter estimate. A variation independent parametrization of an influence curve need not exist, and even if it does exist it can be difficult to identify. For instance, even for estimating the relative risk in a randomized experiment or the cumulative hazard function in a survival analysis problem, it is unclear how estimating function methodology would proceed. For these

two parameters, we recommend performing empirical efficiency maximization through the substitution approach of Section 5.1.

5.4 Multivariate Parameters

A drawback of empirical efficiency maximization is that there is no straightforward generalization to estimating multivariate parameters. It is not necessarily clear that within a misspecified full data working model $\mathcal{F}_0$, there will be a single F making $\frac{1}{n} \sum_{i=1}^n [D_1(O_i|G_0, Q(F)), \dots, D_k(O_i|G_0, Q(F))]^T$ as efficient as possible, meaning this is the best $F \in \mathcal{F}_0$ for approximating any linear combination of $[\mu_1, \dots, \mu_k]^T$.

Of course, multivariate parameter estimation can be handled by breaking the problem into k univariate pieces. While often applicable, such a technique could give unorthodox answers when parameters are known to obey certain orderings, such as when estimating a survival curve at k time points.

Another approach would be to represent the influence curve of multivariate parameter estimate $\hat{\mu} = [\hat{\mu}_1, \dots, \hat{\mu}_k]^T$ as the vector

$$IC(O|G_0, Q, \mu(F_0)) = [IC_1(O|G_0, Q_1, \mu(F_0)), \dots, IC_k(O|G_0, Q_k, \mu(F_0))]^T,$$

for $Q = (Q_1, \dots, Q_k)$. With a consistent preliminary parameter estimate $\hat{\mu}$, the asymptotic covariance matrix (of the estimator applied with known coarsening mechanism) could be approximated empirically with

$$\hat{\Sigma}(Q_1, \dots, Q_k) = \frac{1}{n} \sum_{i=1}^n IC(O_i|\hat{G}, Q, \hat{\mu}) IC(O_i|\hat{G}, Q, \hat{\mu})^T.$$

Empirical efficiency maximization could then operate by defining a norm $\|\cdot\|$ on covariance matrices, and trying to minimize $\|\hat{\Sigma}(Q_1, \dots, Q_k)\|$ over a working index set induced by a full data working model. For many problems, this could be easier said than done.

5.5 Enhancing Model Selection for $\hat{Q}$

Suppose that minimizing empirical risk $\hat{R}(Q) = \frac{1}{n} \sum_{i=1}^n D^2(O_i|\hat{G}, Q)$ over a working index set $\mathcal{Q}$ is difficult, but someone has built candidate fits $\hat{Q}_1, \dots, \hat{Q}_k$ through different procedures. Supposing $Q_1, \dots, Q_k$ are the limits, we could consider performing empirical efficiency maximization with $\mathcal{Q} = \{Q_1, \dots, Q_k\}$, by minimizing $\hat{R}(Q) = \frac{1}{n} \sum_{i=1}^n D^2(O_i|\hat{G}, Q)$ over $\hat{\mathcal{Q}} = \{\hat{Q}_1, \dots, \hat{Q}_k\}$. This method would attempt to select the candidate fit from $\hat{\mathcal{Q}}$ leading to the most

precise parameter estimate. For example, suppose we wish to form covariate-adjusted estimates of five-year survival through fitting a proportional hazards model as in Section 4, and have built models using a variety of different covariate transformations. Instead of finding the transformation maximizing partial likelihood, we would be attempting to target the transformation to best estimate the parameter of interest. Likewise, when forming a fit $\hat{Q}$ in the first place, one could use empirical risk function $\hat{R}(\cdot)$ to select tuning parameters.

5.6 High Dimensional Covariates

We have sought to minimize a risk function $R(Q) = E[D^2(O|G_0, Q)]$, corresponding to the asymptotic variance of a parameter estimate built from nuisance function Q . We have been attempting to minimize risk over a working index set $\mathcal{Q}$, thought of as being generated from a working parametric or semiparametric model for the data generating distribution. When the working index set was relatively small, we considered approximating the unknown risk with empirical risk $\frac{1}{n} \sum_{i=1}^n D^2(O_i|\hat{G}, Q)$.

However, an increasing trend in science is the collection of more and more covariate information. When covariate vector W is extremely high dimensional, it might not make sense to think of building a working model for a conditional distribution given W , such as a logistic regression model or a Cox proportional hazards model. Instead, it can be more appropriate to simply think of $\mathcal{Q}$ as a function class.

It is known in the machine learning literature then when attempting to minimize risk over a large function class, minimizing empirical risk can lead to overfitting. Alternative approaches include regularization, penalization, cross-validation, or local learning. These ideas can be incorporated into our covariate adjustment proposals. For instance, we could consider minimizing $\frac{1}{n} \sum_{i=1}^n D^2(O_i|\hat{G}, Q) + \lambda J(Q)$, for $J(\cdot)$ a penalty functional defined on $\mathcal{Q}$ to penalize complexity, and selecting tuning parameter λ with cross-validation.

The goal would remain ensuring a $\hat{Q} \rightarrow Q$ convergence for which $R(Q) = E[D^2(O|G_0, Q)]$ was small, but it is important to keep in mind that the loss function $L(O, Q|G_0) = D^2(O|G_0, Q)$ can be used more imaginatively than simply through minimizing empirical risk.

5.7 Inference

We have discussed estimators $\hat{\mu}$ for which $\sqrt{n}(\hat{\mu} - \mu) \rightarrow_{\mathcal{L}} N(0, \sigma^2)$, and hence asymptotically valid inference requires a consistent estimate of the asymptotic variance σ^2 . Because $\sigma^2 = E[D^2(O|G_0, Q)] - \mu^2 = R(Q) - \mu^2$, a natural choice

would seem to be $\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n D^2(O_i | \hat{G}, \hat{Q}) - \hat{\mu}^2 = \hat{R}(\hat{Q}) - \hat{\mu}^2$. However, this could be inappropriate. When the coarsening mechanism is efficiently estimated from a correct model, $\hat{\sigma}^2$ will converge to the asymptotic variance of the estimator applied with G_0 known, which will be an overestimate. Unfortunately, the method could also underestimate the variance of $\sqrt{n}(\hat{\mu} - \mu)$ from an overfitting bias, analogous to empirical mean squared error underestimating true mean squared error in regression, although this problem should diminish with sample size. An alternative route would be to construct a variance estimate or confidence interval through the bootstrap.

6 Discussion

One argument for our new estimators may appear circular or flawed. We noted asymptotically efficient estimators are often eschewed because their $\sqrt{n}$ -asymptotics wouldn't encapsulate performance with real data, yet our analysis of estimators was based entirely on $\sqrt{n}$ -asymptotics. Our intuition was that such asymptotics could guide performance with moderately sized samples, when performing adjustment with a sufficiently constrained working model.

We should note that even when the working index set has only two elements, so $\mathcal{Q} = \{Q_1, Q_2\}$, our approach may lead to misbehavior under pathological conditions. If $R(Q_1) = E[D^2(O|G_0, Q_1)] = E[D^2(O|G_0, Q_2)] = R(Q_2)$, then $\hat{Q}$ may oscillate between Q_1 and Q_2 , and it seems likely that examples can be constructed where $\sqrt{n}(\hat{\mu} - \mu)$ does not converge to a Gaussian distribution. As discussed in Section 3.6, a $\hat{Q} \rightarrow Q$ convergence is generally needed, or more precisely a $D(\cdot|G_0, \hat{Q}) \rightarrow D(\cdot|G_0, Q)$ convergence, and can require special care or assumptions. The issue could possibly arise in situations where a working full data model leads to a working index set $\mathcal{Q}$ consisting of nonlinear functions, and the loss function $L(O, Q|G_0) = D^2(O|G_0, Q)$ is not strictly convex in Q . When performing adjustment through standard working models, a potentially important open problem is whether our proposal can be modified to ensure a $\hat{Q} \rightarrow Q$ convergence, and hence asymptotic linearity for the parameter estimate.

Overall, we believe empirical efficiency maximization appears promising. In coarsened data problems where the coarsening mechanism is well understood, our method enhances locally efficient estimation by targeting the full data working model fit for the parameter of interest, and provides a new approach for covariate adjustment in randomized experiments and survival analysis.

Appendix: Proofs of Theorems 1 and 2

Proof of Theorem 1. Note that $\Delta^2 = \Delta$, and that $E[Y\Delta|W] = E[Y_T\Delta|W] = E[Y_T|W]\pi_0(W)$ as Y_T and Δ are independent given W . We observe that,

$$\begin{aligned}\text{Var}(D_T(O|\pi_0, Q_T)) &= E[D_T^2(O|\pi_0, Q_T)] - \mu_T^2 \\ &= E\left[\left|\frac{\Delta Y}{\pi_0(W)}\right|^2\right] + 2R_1 + R_2 - \mu_T^2,\end{aligned}\quad (17)$$

for

$$\begin{aligned}2R_1 &= 2E\left[\frac{\Delta Y}{\pi_0(W)}\left(1 - \frac{\Delta}{\pi_0(W)}\right)Q_T(W)\right] \\ &= 2E\left[Q_T(W)\left(\frac{1}{\pi_0(W)} - \frac{1}{\pi_0^2(W)}\right)\Delta Y\right] \\ &= 2E\left[Q_T(W)\left(\frac{1}{\pi_0(W)} - \frac{1}{\pi_0^2(W)}\right)E[\Delta Y|W]\right] \\ &= 2E\left[Q_T(W)\left(\frac{1}{\pi_0(W)} - \frac{1}{\pi_0^2(W)}\right)E[Y_T|W]\pi_0(W)\right] \\ &= -2E\left[\frac{1 - \pi_0(W)}{\pi_0(W)}E[Y_T|W]Q_T(W)\right],\end{aligned}\quad (18)$$

and

$$\begin{aligned}R_2 &= E\left[\left(1 - \frac{\Delta}{\pi_0(W)}\right)^2 Q_T^2(W)\right] \\ &= E\left[\left(1 - 2\frac{\Delta}{\pi_0^2(W)} + \frac{\Delta}{\pi_0^2(W)}\right)Q_T^2(W)\right] \\ &= E\left[Q_T^2(W)\left(1 - 2\frac{P(\Delta = 1|W)}{\pi_0(W)} + \frac{P(\Delta = 1|W)}{\pi_0^2(W)}\right)\right] \\ &= E\left[Q_T^2(W)\left(1 - 2 + \frac{1}{\pi_0(W)}\right)\right] \\ &= E\left[\frac{1 - \pi_0(W)}{\pi_0(W)}Q_T^2(W)\right].\end{aligned}\quad (19)$$

Combining (17), (18), and (19) we obtain,

$$\begin{aligned}\text{Var}(D_T(O|\pi_0, Q_T)) &= \\ &= -\mu_T^2 + E\left[\left|\frac{\Delta Y}{\pi_0(W)}\right|^2\right] + E\left[\frac{1 - \pi_0(W)}{\pi_0(W)}(Q_T^2(W) - 2E[Y_T|W]Q(W))\right].\end{aligned}$$

The desired result follows after completing the square for $|E[Y_T|W] - Q_T(W)|^2$, and noting that $E[|E[Y_T|W] - Q_T(W)|^2|W] = E[|Y_T - Q_T(W)|^2|W] - E[|Y_T - E[Y|W]|^2|W]$. The final step is observing $E[\frac{1-\pi_0(W)}{\pi_0(W)}|Y_T - Q_T(W)|^2]$ equals $E[\Delta \frac{1-\pi_0(W)}{\pi_0^2(W)}|Y - Q_T(W)|^2]$, as it is easy to check that $E[\frac{\Delta}{\pi_0(W)}\psi(W, Y)] = E[\psi(W, Y_T)]$ for any integrable $\psi(W, Y_T)$. $\square$

Proof of Theorem 2. We first note that the counting process $N(\cdot)$ only jumps when $\Delta = 0$, and that $\Delta(1 - \Delta) = 0$. Hence,

$$\begin{aligned}
 & E \left[\left(\frac{\Delta\psi(\tilde{T}, W)}{G_0(\tilde{T})} \right) \left(\int \frac{Q(u, W)}{G_0(u)} dM(u|G_0) \right) \right] \\
 &= E \left[\left(\frac{\Delta\psi(\tilde{T}, W)}{G_0(\tilde{T})} \right) \left(\int \frac{Q(u, W)}{G_0(u)} (dN(u) - dA(u|G_0)) \right) \right] \\
 &= -E \left[\left(\frac{\Delta\psi(\tilde{T}, W)}{G_0(\tilde{T})} \right) \left(\int \frac{Q(u, W)}{G_0(u)} dA(u|G_0) \right) \right] \\
 &= -E \left[\left(\frac{\Delta\psi(\tilde{T}, W)}{G_0(\tilde{T})} \right) \left(\int \frac{Q(u, W)I(\tilde{T} \geq u)}{G_0^2(u)} dG_0(u) \right) \right] \\
 &= -E \left[\int \frac{I(\tilde{T} \geq u)}{G_0^3(u)} \left(\frac{\Delta\psi(\tilde{T}, W)G_0(u)}{G_0(\tilde{T})} \right) Q(u, W) dG_0(u) \right] \\
 &= -E_{(O, U) \sim \mathcal{L}(O) \times G_0} \left[\left(\frac{I(\tilde{T} \geq U)}{G_0^3(u)} \right) \left(\frac{\Delta\psi(\tilde{T}, W)G_0(U)}{G_0(\tilde{T})} \right) Q(U, W) \right] \quad (20)
 \end{aligned}$$

Further, standard martingale results such as Theorem 2.6.1 in Fleming and Harrington (1991) imply,

$$\begin{aligned}
 & E \left[\int \frac{Q(u, W)}{G_0(u)} dM(u|G_0) \right]^2 = E \int \frac{Q^2(u, W)}{G_0^2(u)} dA(u|G_0) \\
 &= E \int \left(\frac{I(\tilde{T} \geq u)}{G_0^3(u)} \right) Q^2(u, W) dG_0(u) \\
 &= E_{(O, U) \sim \mathcal{L}(O) \times U} \left[\left(\frac{I(\tilde{T} \geq U)}{G_0^3(U)} \right) Q^2(U, W) \right]. \quad (21)
 \end{aligned}$$

The result follows after squaring the sum $D(O|G_0, Q)$, finding the expectation of the terms with (20) and (21), and completing the square for the quadratic in $Q(U, W)$. $\square$

References

- [1] Alonzo, T.A., Pepe, M.S., and Lumley, T.S. (2003). Estimating disease prevalence in two-phase studies. *Biostatistics*. 4, 2, 313-326.
- [2] Bickel, P.J., Klaassen, C.A.J., Ritov, Y., and Wellner, J.A (1998). *Efficient and Adaptive Estimation for Semiparametric Models*. Springer-Verlag, New York.
- [3] Breslow, N.E. (1974). Covariance analysis of censored survival data. *Biometrics*. 30, 89-99.
- [4] Clayton, D., Dunn, G., Pickles, A., and Spiegelhalter, D. (1998). Analysis of longitudinal binary data from multiphase sampling. *Journal of the Royal Statistical Society, Series B*, 60, 71-87.
- [5] Cox, D.R. (1972). Regression models and life tables (with discussion). *Journal of the Royal Statistical Society B*, 34, 187-220.
- [6] Fisher, R.A. (1932). *Statistical Methods for Research Workers*. Edinburgh: Oliver and Boyd (13th ed., 1958).
- [7] Fleming, T.R. and Harrington, D.P. (1991). *Counting Processes and Survival Analysis*. John Wiley & Sons, Inc.
- [8] Freedman, D.A. (2007). On regression adjustments to experimental data. To appear in *Advances in Applied Mathematics*. Available at <http://www.stat.berkeley.edu/census/neyregr.pdf>
- [9] Freedman, D.A. (2008). Randomization does not justify logistic regression. Available at <http://www.stat.berkeley.edu/census/neylogit.pdf>
- [10] Friedman, L.M., Furberg, C.D., and DeMets, D.L. (1998). *Fundamentals of Clinical Trials* (third edition). Springer-Verlag, New York.
- [11] Gill, R.D., van der Laan, M.J., and Robins, J.M. (1997). Coarsening at random: characterizations, conjectures and counter-examples. *Proceedings of the First Seattle Symposium in Biostatistics*, 1995. Springer Lecture Notes in Statistics, 255-294.
- [12] Heitjan, D.F. and Rubin, D.B. (1991). Ignorability and coarse data. *Annals of Statistics*, 19, 2244-2253.

- [13] Horvitz, D.G. and Thompson, D.J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*. 47, 663-685.
- [14] Kaplan, E.L., and P. Meier. (1958). Nonparametric estimator from incomplete observations. *Journal of the American Statistical Association*. 53, 457-481.
- [15] Koch GG, Tangen CM, Jung JW, Amara IA (1998). Issues for covariance analysis of dichotomous and ordered categorical data from randomized clinical trials and non-parametric strategies for addressing them. *Statistics in Medicine*, 17, 1863-1892.
- [16] Lu, X. *Improving the efficiency of the log-rank test using auxiliary covariates*. Presentation on August 10, 2006 in Seattle, WA at the Joint Statistical Meeting of the American Statistical Association.
- [17] 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.
- [18] Neyman, J. (1923) Sur les applications de la théorie des probabilités aux expériences 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).
- [19] Neyman, J. (1938). Contribution to the theory of sampling human populations. *Journal of the American Statistical Association*. 33, 101-116.
- [20] 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.
- [21] 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.
- [22] 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.

- [23] Robins J.M. and Rotnitzky, A. (2005). Inverse probability weighted estimation in survival analysis. *Encyclopedia of Biostatistics*, Second Edition, Editors: Armitage, P. and Colton, T., Wiley & Sons, New York.
- [24] Robinson, L.D. and Jewell, N.P. (1991). Some surprising results about covariate adjustment in logistic regression. *International Statistical Review*, 59, 227-240.
- [25] Rubin, D.B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66, 688-701.
- [26] Rubin, D.B. and van der Laan, M.J. (2008). Covariate Adjustment for the Intention-to-Treat Parameter with Empirical Efficiency Maximization." *U.C. Berkeley Division of Biostatistics Working Paper Series*. Working Paper 229.
- [27] Satten, G.A, Datta, S., and Robins, J.M. (2001). Estimating the marginal survival function in the presence of time dependent covariates. *Statistics & Probability Letters*. 54, 4, 397-403.
- [28] Scharfstein, D.O. and Robins, J.M. (2002). Estimation of the failure time distribution in the presence of informative censoring. *Biometrika*.. 89(3), 617-634.
- [29] 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.
- [30] Tsiatis, A.A. (2006). *Semiparametric Theory and Missing Data*. Springer Science + Business Media, LLC.
- [31] Tsiatis, A.A., Davidian, M., Zhang, M., and Lu, X. (2007). Covariate adjustment for two-sample treatment comparisons in randomized clinical trials: A principled yet flexible approach. *Statistics in Medicine*.
- [32] van der Laan, M.J. and Robins, J.M. (2003). *Unified Methods for Censored Longitudinal Data and Causality*. Springer-Verlag, New York.
- [33] Zeng, D. (2004). Estimating marginal survival function by adjusting for dependent censoring using many covariates. *Annals of Statistics*. Vol. 32, No. 4, 1533-1555.