University of California, Berkeley

U.C. Berkeley Division of Biostatistics Working Paper Series

Year 2007 *Paper* 220

Empirical Efficiency Maximization

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

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

†Division of Biostatistics, School of Public Health, University of California, Berkeley, laan@berkeley.edu

Empirical Efficiency Maximization

Daniel B. Rubin and Mark J. van der Laan

Abstract

It has long been recognized that covariate adjustment can increase precision, even when it is not strictly necessary. The phenomenon is particularly emphasized in clinical trials, whether using continuous, categorical, or censored time-to-event outcomes. Adjustment is often straightforward when a discrete covariate partitions the sample into a handful of strata, but becomes more involved when modern studies collect copious amounts of baseline information on each subject.

The dilemma helped motivate locally efficient estimation for coarsened data structures, as surveyed in the books of van der Laan and Robins (2003) and Tsiatis (2006). Here one fits a relatively small working model for the full data distribution, 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 is asymptotically efficient if the working model is correct, but otherwise is still consistent and asymptotically Normal.

However, the working model will almost always be misspecified in practice. By applying standard likelihood based fits, one can poorly estimate the parameter of interest. We propose a new method, empirical efficiency maximization, to target the element of a working model minimizing asymptotic variance for the resulting parameter estimate, whether or not the working model is correctly specified.

Our procedure is illustrated in three examples. It is shown to be a potentially major improvement over existing covariate adjustment methods for estimating disease prevalence in two-phase epidemiological studies, treatment effects in two-arm randomized trials, and marginal survival curves. Numerical asymptotic efficiency calculations demonstrate gains relative to standard locally efficient estimators.

1 Introduction

Consider an experiment in which a covariate vector $W \in \mathbb{R}^p$ is measured on a subject at baseline, the subject is randomly assigned to a treatment group with probability π_0 or a control group with probability $1-\pi_0$, and an outcome $Y \in \mathbb{R}$ is then assessed. The observed data would consist of n replicates $\{W_i, \Delta_i, Y_i\}_{i=1}^n$, for $\Delta_i \in \{0,1\}$ a treatment indicator. Suppose clinical interest lies in assessing a treatment effect

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

For a binary outcome Y, such as a disease indicator, parameter (1) is termed the excess risk. Treatment effects could also correspond to (log) relative risks or (log) odds ratios between disease probabilities $P(Y = 1|\Delta = 1)$ and $P(Y = 1|\Delta = 0)$.

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

$$\mu_n = \frac{\sum_{i=1}^n \Delta_i Y_i}{\sum_{i=1}^n \Delta_i} - \frac{\sum_{i=1}^n (1-\Delta_i) Y_i}{\sum_{i=1}^n 1-\Delta_i}, \tag{2}$$

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

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 the treatment and control arm, while unadjusted analysis cannot exploit such knowledge .

The semiparametric literature (e.g. Bickel et al., 1998) has traditionally focused on efficient estimators. Unfortunately, such estimators suffer from the curse of dimensionality in nontrivial covariate adjustment problems, and the asymptotically efficient estimator of treatment effect μ in (1) would involve consistently estimating the function $Q(w, \delta) = E[Y|W = w, \Delta = \delta]$, and using the fit 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 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 fits a relatively small working model for the data generating distribution. For a binary outcome Y, this might take the form of a logistic regression model. One then takes Q_n to be the plug-in estimator of $Q : (w, \delta) \to E[Y|W = w, \Delta = \delta]$, and uses this Q_n as a nuisance parameter fit in the efficient estimating equation for the treatment effect. Asymptotic efficiency is achieved if the working model holds, while a misspecified working model does not compromise $\sqrt{n}$-consistency or asymptotic Normality. 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 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. When the working model is incorrect, such a fit can be a very poor choice for the resulting treatment effect estimate. In fact, we will see that 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 knowing how to better use working models for covariate adjustment, little will be lost by skipping ahead. Section 3 considers prevalence estimation in two-phase studies. We return to treatment effect estimation in Section 4, and the omnipresent marginal survival problem is examined in Section 5. Numerical asymptotic efficiency calculations demonstrate advantages over existing methods. Sections 6 and 7 discuss extensions and comparisons with other covariate adjustment procedures, and an appendix provides templates for proving empirical efficiency maximization leads to asymptotically linear estimators.

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 $\mathcal{F}_0$ is a statistical model. Suppose our interest would be in estimating a smooth full data parameter $\mu(F_0) \in \mathbb{R}^k$. For the time being, we will restrict attention to estimating a population mean $\mu(F_0) = E_{F_0}[\psi(X)] \in \mathbb{R}$, and defer treatment of general smooth parameters to Section 6.

If we could observe the full data, we could estimate $\mu(F_0)$ with the empirical mean $\frac{1}{n}\sum_{i=1}^n \psi(X_i)$. When $\psi(X)$ has finite variance and the full data model $\mathcal{F}$ is nonparametric, the empirical mean is asymptotically efficient.

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 of $\{C|X\}$ is known as the coarsening mechanism, belonging to the model $\mathcal{G}$. We'll assume 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 $\mu(F_0)$ 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 from this submodel we can efficiently estimate G_0 with G_n.

An estimator sequence $\mu_n = \mu_n(O_1, ..., O_n)$ is said to be asymptotically linear with influence curve $IC(O|P_0) \in L_0^2(P_0)$ if

$$\mu_n = \mu(F_0) + \frac{1}{n}\sum_{i=1}^{n} IC(O_i|P_0) + o_{P_0}(n^{-1/2}). \tag{3}$$

For a population mean $\mu(F_0) = E_{F_0}[\psi(X)]$, the efficient estimator's influence curve takes the form

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

where $D(O|G_0, Q(F_0))$ results from the doubly robust mapping of $\psi(X)$, defined in Theorem 2.1 of van der Laan and Robins (2003). An estimator μ_n satisfying (3) will be asymptotically Normal, meaning $\sqrt{n}(\mu_n - \mu(F_0)) \rightarrow_{\mathcal{L}} N(0, \sigma^2)$. Efficiency at P_0 implies that if there is another regular estimator sequence $\{\hat{\mu}_n\}$ for which $\sqrt{n}(\hat{\mu}_n - \mu(F_0)) \rightarrow_{\mathcal{L}} N(0, \tau^2)$, then $\sigma^2 \leq \tau^2$, so the sequence estimates parameter $\mu(F_0)$ with less precision. The efficient estimator's influence curve is termed the efficient influence curve. If $\mu_n - \hat{\mu}_n = o_{P_0}(n^{-1/2})$, the estimators are said to be asymptotically equivalent.

Equations (3) and (4) clearly suggest a route to efficient estimation of $\mu(F_0)$: fit the nuisance parameter $Q(F_0)$ from the data with Q_n, and then apply $\mu_n = \frac{1}{n}\sum_{i=1}^{n} D(O_i|G_n, Q_n)$. 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, estimating the nuisance parameter $Q(F_0)$ often requires solving a high dimensional function approximation problem. Robins and Ritov (1997) provide an extended discussion, 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

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

Locally efficient coarsened data estimators have operated by letting Q_n be the efficient estimate of nuisance parameter $Q(F_0)$ under working model $\mathcal{F}_0$. The estimator $\frac{1}{n}\sum_{i=1}^{n} D(O_i|G_n, Q_n)$ 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

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

meaning a misspecified nuisance parameter $Q(F) \neq Q(F_0)$ will not compromise the estimator. A salient feature is the double protection property that if either the working model $\mathcal{F}_0$ or the coarsening mechanism model $\mathcal{G}_0$ is correctly specified, asymptotic linearity can be achieved. Section 7 contrasts doubly robust estimation with our forthcoming proposal.

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 $\mathcal{Q}_0$ is often simply an index of estimators, excluding many $Q \in \mathcal{Q}$ 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 $\mu(F_0)$.

One way to gauge the quality of $Q \in \mathcal{Q}_0$ is with the asymptotic variance of $\mu_{n,Q} = \frac{1}{n}\sum_{i=1}^{n} D(O_i|G_0, Q)$, which by the robustness result (5) is given by

$$\sigma^2(Q) = \mathrm{Var}_{P_0}(D(O|G_0, Q)) = E_{P_0}[D^2(O|G_0, Q)] - \mu^2(F_0). \tag{6}$$

When the full data working model fit Q_n converges to an element $Q \in \mathcal{Q}_0$, the asymptotic variance of our estimator $\frac{1}{n}\sum_{i=1}^{n} D(O_i|G_0, Q_n)$ will be $\sigma^2(Q)$, and hence it is vital to consider our fit's limiting behavior.

Note that $\mu_{n,Q}$ is unbiased for the parameter $\mu(F_0)$, but can only be applied as an estimator with known coarsening mechanism. When our estimator is $\frac{1}{n}\sum_{i=1}^{n} D(O_i|G_n, Q)$, and G_n 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. We discuss this further in the appendix, 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 $Q \in \mathcal{Q}_0$ 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 of $E_{P_0}[D^2(O|G_0, Q)]$, and we can estimate this population mean with the empirical mean $\frac{1}{n}\sum_{i=1}^{n} D^2(O_i|G_n, Q)$. The approximation should be valid for all $Q \in \mathcal{Q}_0$, irrespective of whether this working index set was induced by a correctly specified working model for the data generating distribution. If $\mathcal{Q}_0$ 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 uniformly be close to each other, and for the empirical minimizer to approach the population minimizer. We therefore propose selecting

$$Q_n = \mathrm{argmin}_{Q \in \mathcal{Q}_0} \frac{1}{n}\sum_{i=1}^{n} D^2(O_i|G_n, Q),$$

and applying the estimator $\mu_n = \frac{1}{n}\sum_{i=1}^{n} D(O_i|G_n, Q_n)$.

There can in fact be several minimizers of $\sigma^2(Q)$ over working index set $\mathcal{Q}_0$, and what we precisely mean is finding the function in

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

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. While we assume our working model leads to an optimal element in $\mathcal{D}_0$, we will throughout this work refer to finding optimal elements of $\mathcal{Q}_0$ or $\mathcal{F}_0$.

Under regularity conditions discussed in the appendix, our estimator $\sqrt{n}(\mu_n - \mu(F_0))$ will converge in law to a Normal distribution, with variance no larger than the infimum of $\sigma^2(Q)$ as Q ranges over the working index set $\mathcal{Q}_0$. Two consequences are of note:

1. Our estimator is locally efficient, so will be asymptotically equivalent to standard locally efficient estimators if the working model $\mathcal{F}_0$ 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 the unadjusted estimator ignoring covariates is asymptotically equivalent to $\frac{1}{n}\sum_{i=1}^{n} D(O_i|G_0, Q)$ for some $Q \in \mathcal{Q}_0$. Our estimator is then guaranteed to be at least as efficient as the unadjusted techniques. This property is not shared by prevailing locally efficient estimators, as we will see in two examples.

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 $\mathcal{Q}_0$ to minimize a risk function $R : \mathcal{Q} \to \mathbb{R}$, 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 nuisance parameter G_0, because it is associated with risk $\sigma^2(Q)$. Hence, selecting $Q \in \mathcal{Q}_0$ is an intermediate step in applying the estimator $\mu_n = \frac{1}{n}\sum_{i=1}^{n} D(O_i|G_n, Q)$, and we estimate $Q_n \in \mathcal{Q}_0$ 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 Two-Phase Designs

Following their introduction by 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.

Formally, we can let the vector $W \in \mathbb{R}^p$ denote information collected on a subject in phase 1, let $\Delta \in \{0, 1\}$ be an indicator of phase 2 inclusion, and let Y be an indicator

of disease status. A subject's observed data is then $O \sim P_0$, for

$$O = (W, \Delta, \Delta Y).$$

The subject's potentially unobserved full data is $X = (W, Y) \sim F_0$, meaning we would have liked to use the gold standard diagnostic technique for everyone in the sample.

We will see that efficient estimation is straightforward if the phase 1 measurement W takes on only a handful of values to determine phase 2 sampling strata, such as when it only consists of a "cheap and easy-to-use surrogate disease indicator." When demographic and clinical measurements are also collected on all subjects, prevalence estimation can be made with better precision, but requires more thought. Hence, we can consider estimation in two-phase studies as a covariate adjustment problem, where covariates represent phase 1 information.

We slightly modify standard analysis, by considering observing $\{O_i\}_{i=1}^n$ i.i.d., meaning that each subject is included in the second phase based on a weighted coin flip, where weights are determined by phase 1 information, but the coin flip doesn't depend on other subjects' coin flips. Studies would more likely have a fixed phase 2 sample size in mind. There can be subtle differences in an estimator's asymptotics in the two designs, which will have to be finessed for empirical efficiency maximization at a later date. Under mild conditions, we expect in both situations to make the optimal locally efficient covariate adjustment from a full data working model.

We let $\pi_0(W) = P_{G_0}(\Delta = 1|W)$ denote the probability of phase 2 inclusion given covariates W observed in phase 1. Note that $\pi_0(\cdot)$ is assumed to be a known function, and plays the role of coarsening mechanism G_0 previously discussed. In this case, we do not have to estimate $\pi(\cdot)$ with $\pi_n(\cdot)$ according to a correctly specified model, although our final estimator would be no less efficient. For identifiability, we assume $\pi_0(W)$ is bounded away from zero with probability one. As phase 2 inclusion is randomly determined following phase 1 measurements, we also assume the conditional independence $\{\Delta \perp Y|W\}$, implying coarsening at random. The parameter of interest is defined to be $\mu(F_0) = E_{F_0}[Y]$. When $Y \in \{0, 1\}$ is a disease indicator, the disease prevalence is $\mu(F_0) = E_{F_0}[Y] = P_{F_0}(Y = 1)$.

3.1 Estimation

A popular approach is to use the (1952) Horvitz-Thompson estimator

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

averaging phase 2 responses, but weighing by the inverse of phase 1 inclusion probabilities. While simple and unbiased, the estimator can be quite inefficient, because it ignores information collected from those subjects only assessed in phase 1.

The efficient influence curve for $\mu(F_0)$ is given in van der Laan and Robins (2003) as

$$IC(O|\pi_0, Q(F_0), \mu(F_0)) = D(O|\pi_0, Q(F_0)) - \mu(F_0),$$

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

$$D(O|\pi_0, Q(F_0)) = \frac{\Delta Y}{\pi_0(W)} + (1 - \frac{\Delta}{\pi_0(W)})Q(F_0)(W),$$

where

$$Q(F) : W \to E_F[Y|W] = P_F(Y = 1|W)$$

maps full data distributions $F \in \mathcal{F}$ to functions from $\mathcal{W}$ to $\mathbb{R}$.

In line with robustness result (5), the estimator

$$\mu_{n,Q} = \frac{1}{n}\sum_{i=1}^{n} D(O_i|\pi_0, Q)$$

will be unbiased for $\mu(F_0) = P_{F_0}(Y = 1)$ at any $Q(F) : W \to P_F(Y = 1|W)$. This can be immediately seen, because it is a trivial computation to show the second term in $D(O|\pi_0, Q)$ has mean zero for any Q, while the first term has expectation $\mu(F_0)$ as in the Horvitz-Thompson estimator. Each $Q \in \mathcal{Q} = \{Q(F) : F \in \mathcal{F}\}$ leads to an asymptotic variance $\sigma^2(Q)$ of estimator $\mu_{n,Q}$. The choice $Q(F_0)(W) = P_{F_0}(Y = 1|W)$ minimizes this quantity.

An asymptotically efficient estimator can be constructed for this problem, and would involve consistently fitting the binary regression $Q(F_0)(W) = P_{F_0}(Y = 1|W)$ with Q_n, and then applying $\mu_{n,Q_n} = \frac{1}{n}\sum_{i=1}^{n} D(O_i|\pi_0, Q_n)$. From our assumed conditional independence $\{\Delta \perp Y|W\}$, it follows that $P_{F_0}(Y = 1|W) = P_{F_0}(Y = 1|W, \Delta = 1)$, so we could fit the regression function using only the phase 2 sample. Note that μ_{n,Q_n} would no longer necessarily be unbiased for $\mu(F_0)$.

Efficiency presents no essential challenge when W takes on a small number of values, because $Q(F_0)(w)$ could be fit with the empirical mean of Y in the $\{W = w\}$ stratum of phase 2. For W a large vector of informative covariates measured in phase 1, consistently fitting $P_{F_0}(Y = 1|W)$ would be a potentially difficult function approximation problem, susceptible to the curse of dimensionality. When simple unbiased estimators such as the Horvitz-Thompson weighted average are available, data analysts might be justifiably hesitant to apply machine learning tools.

The locally efficient technique is to assume a relatively constrained full data working model $\mathcal{F}_0 \subset \mathcal{F}$, inducing a working index set

$$\mathcal{Q}_0 = \{Q(F) : W \to P_F(Y = 1|W) : F \in \mathcal{F}_0\} \subset \mathcal{Q}. \tag{7}$$

Locally efficient procedures have traditionally fit full data working models with maximum likelihood, and such fits were mentioned for prevalence estimation in two-phase studies by Clayton et al. (1998) and Alonzo, Pepe, and Lumley (2003). A correctly specified working model leads to asymptotic efficiency. The estimator μ_{n,Q_n} will be asymptotically linear even when Q_n is fit from a misspecified working model, and the hope is to still gain precision by making some use of informative phase 1 information.

By far the most popular working model for the conditional distribution of a binary variable is the logistic regression model, in which case our working index set would be

$$\mathcal{Q}_0 = \{Q_\beta(w) \to \frac{1}{1 + \exp(-\beta^T w)} : \beta \in \mathbb{R}^p\},$$

where our notation suppresses the intercept by including a constant element in the W vector. Note that locally efficient estimation differs from the plug-in estimator $\mu_n = \frac{1}{n}\sum_{i=1}^n Q_n(W_i)$ resulting from the logistic regression fit of $P_{F_0}(Y=1|W)$ and the empirical distribution for $\mathcal{L}(W)$. The plug-in technique isn't robust against working model misspecification, and consequently is inconsistent.

When the logistic regression model is correctly specified, the usual maximum likelihood estimate of coefficient vector β will of course converge to the true coefficient vector. But when the model is incorrect, the maximum likelihood fit will converge to some $Q \in \mathcal{Q}_0$ (that minimizing Kullback-Leibler divergence from $Q(F_0)$), and this might not be the optimal element for estimating the parameter of interest. Empirical efficiency maximization notes that the asymptotic variance $\sigma^2(Q)$ of $\mu_{n,Q}$ can be approximated empirically with $\frac{1}{n}\sum_{i=1}^n D^2(O_i|\pi_0, Q)$, and tries to minimize this empirical "risk" over working index set $\mathcal{Q}_0$. This here reduces to regressing $\frac{\Delta Y}{\pi_0(W)}$ on $(\frac{\Delta}{\pi_0(W)} - 1)Q_\beta(W)$, and solving for β with nonlinear least squares.

In a broader sense, we refer to empirical efficiency maximization as the class of methods that represent $\sigma^2(Q)$ as monotone in the population mean of a function of the observed data, nuisance parameter, and working model element Q, and attempt to minimize the empirical mean of this function over the working index set. In two-phase prevalence estimation, there is another representation of the asymptotic variance we can use for minimization. Examine the following theorem, proven in the appendix.

Theorem 1.

$$\sigma^2(Q) = Var_{P_0}(D(O|\pi_0, Q)) = C(P_0) + E_{P_0}[\Delta\frac{1-\pi_0(W)}{\pi_0^2(W)}|Y - Q(W)|^2]$$

for

$$C(P_0) = -\mu^2 + E[|\frac{\Delta Y}{\pi_0(W)}|^2] + E[\frac{1-\pi_0(W)}{\pi_0(W)}\{E[Y|W]^2 - (Y - E[Y|W])^2\}]$$

not depending on Q.

The theorem reveals that the optimal $Q \in \mathcal{Q}_0$ for efficiency is the binary regressor minimizing not a Kullback-Leibler divergence, but a weighted mean squared error. If the logistic regression does not hold exactly, the two minima might differ. The empirical efficiency maximization approach is thus to fit logistic regression coefficients with weighted nonlinear least squares, which can be done with the **nls()** function in the R language. We must solve

$$\beta_n = \text{argmin}_{\beta \in \mathbb{R}^p} \frac{1}{n}\sum_{i=1}^n \Delta_i \frac{1-\pi_0(W_i)}{\pi_0^2(W_i)}|Y_i - Q_\beta(W_i)|^2,$$

and then estimate our parameter with $\mu_{n,Q_{\beta_n}} = \frac{1}{n}\sum_{i=1}^n D(O_i|\pi_0, Q_{\beta_n})$. The appendix provides templates for proving this estimator is asymptotically equivalent to $\mu_{n,Q_{\beta_0}} = \frac{1}{n}\sum_{i=1}^n D(O_i|\pi_0, Q_{\beta_0})$, where Q_{β_0} is the element of $\mathcal{Q}_0$ optimizing asymptotic variance.

It may initially appear counterintuitive to fit logistic regression coefficients with a criterion other than the likelihood function. It is important to keep in mind that

the stated goal is estimating prevalence $\mu(F_0) = P_{F_0}(Y = 1)$, and that estimating the binary regression function $P_{F_0}(Y = 1|W)$ is not an end in itself. While weighted nonlinear least squares might lead to an unappealing regression fit, it will in general be preferable for our parameter of interest.

When the logistic regression assumptions are satisfied, weighted nonlinear least squares will be less efficient than maximum likelihood for the model coefficients, but equally efficient for prevalence estimate $\frac{1}{n}\sum_{i=1}^{n} D(O_i|\pi_0, Q_n)$. This occurs because as we argue in the appendix, the rate that Q_n converges to the optimal $Q_{\beta_0} \in \mathcal{Q}_0$ will not affect asymptotics, so long as the coarsening mechanism is estimated at rate $n^{-1/2}$, and in this example it is known by design.

Note the Horvitz-Thompson estimator is of the form $\mu_{n,Q} = \frac{1}{n}\sum_{i=1}^{n} D(O_i|\pi_0, Q)$, with constant function $Q(W) = 0$. The logistic regression working index set $\mathcal{Q}_0$ includes all constant functions in $(0, 1)$, and we do not expect discontinuity of $\sigma^2(Q)$ as Q approaches the zero function at index set's boundary. Hence, the optimal working model element $Q_{\beta_0} \in \mathcal{Q}_0$ should improve upon the Horvitz-Thompson estimator's asymptotic variance. When using empirical efficiency maximization to be locally efficient, we therefore might guarantee asymptotic gains over the naive estimator ignoring phase 1 information.

3.2 Asymptotic Variance Calculations

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

$$
\begin{aligned}
&W \sim \mathrm{N}(0, 1)\\
&\pi_0(W) = P_{G_0}(\Delta = 1|W) = \frac{1}{1 + \exp(W)} \text{ truncated to be in } [.1, .9]\\
&P_{F_0}(Y = 1|W) = \frac{1}{1 + \exp(W + \eta W^2)}.
\end{aligned}
$$

Here η determined misspecification of the working logistic regression model, 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 prevalence $\mu(F_0) = P_{F_0}(Y = 1)$. Based on a new sample of size $n = 100,000$, we then computed asymptotic variances by evaluating $\frac{1}{n}\sum_{i=1}^{n} D^2(O_i|\pi_0, Q) - \mu^2(F_0)$. We considered Q 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 $Q(F_0)(W) = P_{F_0}(Y = 1|W)$.

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 phase 1 information. 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.

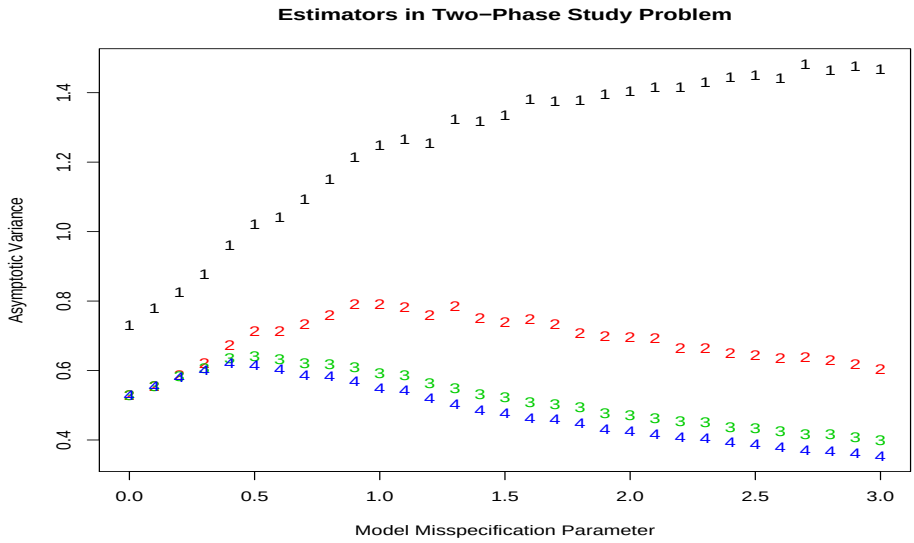

Figure 1: 1−4 represent the Horvitz-Thompson estimator that ignores phase 1 measurements, a locally efficient estimator fitting a misspecified logistic regression model's coefficients with maximum likelihood, our estimator fitting logistic regression coefficients with empirical efficiency maximization, and an asymptotically efficient estimator using empirical mean $\mu_n = \frac{1}{n}\sum_{i=1}^{n} D(O_i|\pi_0, Q)$ at the true $Q(F_0)(W) = P_{F_0}(Y = 1|W)$.

4 Randomized Experiments with Covariates

We now return to the treatment effect estimation problem considered in the introduction, and fill in details for empirical efficiency maximization. Recall that the observed data is $O = (W, \Delta, Y) \sim P_0$, where $W \in \mathbb{R}^p$ is a baseline covariate vector, $\Delta \in \{0, 1\}$ is a treatment indicator, and $Y \in \mathbb{R}$ is an assessed outcome. Subjects are randomly assigned to the treatment group with probability π_0. Often $\pi_0 = \frac{1}{2}$ in clinical trials, and we assume $0 < \pi_0 < 1$ for identifiability.

To cast estimation in our general setting of Section 2, we must consider what unavailable full data X we would have liked to measure for each subject. For this purpose we use the counterfactual outcome formulation proposed by Neyman (1923) and Rubin (1974) (Donald B. Rubin, not the D.B. Rubin authoring the present work). A subject's full data is then $X = (W, (Y_1, Y_0)) \sim F_0 \in \mathcal{F}$, where Y_1 and Y_0 are the outcomes that would have occurred under treatment or no treatment, only one of which is ever seen. The known assignment probability π_0 plays the role of coarsening mechanism G_0 in Section 2. As in (1), we consider estimating treatment effect

$$\mu(F_0) = E_{F_0}[Y_1 - Y_0] = E_{F_0}[Y_1] - E_{F_0}[Y_0] = E_{P_0}[Y|\Delta = 1] - E_{P_0}[Y|\Delta = 0].$$

Before delving into estimation, a few words are in order regarding sampling assumptions. As in the two-phase design problem, we depart from Neyman's original conception and assume an i.i.d. sample $\{O_i\}_{i=1}^n$, meaning each subject is assigned to treatment or control based on a (possibly unfair) coin flip, but subjects' coins don't

influence one another. Friedman, Furberg, and DeMets (1998) note 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 form a subsample of the n subjects to ensure that proportion π_0 were assigned to treatment. Further, randomization ensures the two groups are balanced on important prognostic factors measured at baseline. It has been claimed that this can increase power, but Peto (1978) and Mantel (1984) have argued that proper covariate adjustment leads to equivalent asymptotics under unstratified randomization. Our restriction to an i.i.d. treatment assignment scheme is for exposition. We expect our locally efficient adjustment procedures to have optimal $\sqrt{n}$-asymptotics under more realistic sampling, and intend to show in future work why the coin flipping design results in no precision loss for our estimators.

To test the hypothesis of no difference between treatment and control arms, one can compute the null distribution of a treatment effect statistic by repeating the randomization to assign treatment labels, and repeatedly recomputing the parameter fit.

No matter the sampling scheme, it is usually possible to compute a valid treatment effect estimate by ignoring baseline covariates. The unbiased Horvitz-Thompson analog using inverse probability weighting is easily seen to be

$$\mu_n = \frac{1}{n}\sum_{i=1}^{n}\{\frac{\Delta_i Y_i}{\pi_0} - \frac{(1-\Delta_i)Y_i}{1-\pi_0}\}.$$

We've mentioned that fitting the coarsening mechanism can only improve efficiency. Replacing known π_0 with the empirical $\pi_n = \frac{1}{n}\sum_{i=1}^{n}\Delta_i$ reduces to applying (2), which is asymptotically efficient among unadjusted estimators.

Life becomes less simple with covariates, particularly when too much baseline information is collected to naturally partition the sample. Pocock et al. (2002) surveyed 50 clinical trial reports, and found that 36 used covariate adjustment, and that 12 reports emphasized adjusted over unadjusted analysis. The authors remarked that "Nevertheless, the statistical emphasis on covariate adjustment is quite complex and often poorly understood, and there remains confusion as to what is an appropriate statistical strategy."

4.1 Proposed Treatment Effect Estimator

Efficiency theory can guide the path to appropriate adjustment. van der Laan and Robins (2003) show the efficient influence curve for the treatment effect (1) is

$$IC(O|\pi_0, Q(F_0), \mu(F_0)) = D(O|\pi_0, Q(F_0)) - \mu(F_0). \tag{8}$$

Here $D(O|\pi_0, Q(F_0))$ is the doubly robust mapping of $\psi(X) = Y_1 - Y_0$, and is given by

$$D(O|\pi_0, Q(F_0)) = \{\frac{\Delta Y}{\pi_0} + (1-\frac{\Delta}{\pi_0})Q(F_0)(W,1)\} - \{\frac{(1-\Delta)Y}{1-\pi_0} + (1-\frac{1-\Delta}{1-\pi_0})Q(F_0)(W,0)\},$$

where

$$Q(F)(w,\delta) \to E_F[Y|W=w, \Delta=\delta]$$

maps full data distributions $F \in \mathcal{F}$ to functions from $\mathcal{W} \times \{0,1\}$ to $\mathbb{R}$.

Analogous to the two-phase design problem, efficient estimation would entail consistently estimating the regression function $Q(F_0)$ for use in $\mu_n = \frac{1}{n}\sum_{i=1}^n D(O_i|\pi_0, Q_n)$, and the procedure would be unreliable for nontrivial baseline covariates.

Local efficiency is again based on the implication of robustness result (5) that $\mu_{n,Q} = \frac{1}{n}\sum_{i=1}^n D(O_i|\pi_0, Q(F))$ is unbiased for treatment effect $\mu(F_0)$ at any $Q(F) \in \mathcal{Q} = \{Q(F) : F \in \mathcal{F}\}$. We proceed by using a working model $\mathcal{F}_0$ for the full data distribution, inducing a set of functions $\mathcal{Q}_0 = \{Q(F) : F \in \mathcal{F}_0\}$, and then applying $\mu_n = \frac{1}{n}\sum_{i=1}^n D(O_i|\pi_0, Q_n)$ for appropriate fit $Q_n \in \mathcal{Q}_0$.

For continuous outcome Y, we could model the function $Q(F_0)(w,\delta) = E_{F_0}[Y|W = w, \Delta = \delta]$ with linear regression. The working index set would then become

$$\mathcal{Q}_0 = \{Q_{\beta_0,\beta_1,\beta_2}(w,\delta) = \beta_0 + \beta_1\delta + \beta_2^T w : \beta_0, \beta_1 \in \mathbb{R}, \beta_2 \in \mathbb{R}^p\}.$$

One's first thought might be to use working index set fit Q_n corresponding to the least squares solution

$$(\beta_{0,n}, \beta_{1,n}, \beta_{2,n}) = \text{argmin}_{\beta_0,\beta_1,\beta_2} \frac{1}{n}\sum_{i=1}^n |Y_i - \beta_0 - \beta_1\Delta_i - \beta_2^T W_i|^2,$$

which of course reduces to the maximum likelihood estimate under i.i.d. Gaussian errors in the regression model.

By an algebraic coincidence, the least squares fit Q_n leads to an identical locally efficient estimate $\mu_{n,Q_n} = \frac{1}{n}\sum_{i=1}^n D(O_i|\pi_0, Q_n)$, plug-in estimate $\frac{1}{n}\sum_{i=1}^n \{Q_n(W_i,1) - Q_n(W_i,0)\}$ of $\mu(F_0) = E[E[Y|W,\Delta=1] - E[Y|W,\Delta=0]]$, and coefficient estimate $\beta_{1,n}$. It follows immediately from the theory of local efficiency that this estimate is asymptotically linear, even if the linear regression model is misspecified. Under a slightly altered sampling scheme, Freedman (2007a) proves that "almost anything can happen" when $\pi_0 \neq \frac{1}{2}$, meaning asymptotic variance can be better or worse than that of the efficient unadjusted estimator (2).

Empirical efficiency maximization instead proceeds by attempting to best fit working index set $\mathcal{Q}_0$ for the parameter of interest. We again propose selecting

$$Q_n = \text{argmin}_{Q \in \mathcal{Q}_0} \frac{1}{n}\sum_{i=1}^n D^2(O_i|\pi_0, Q). \tag{9}$$

For the working index set $\mathcal{Q}_0$ induced by a linear regression model, this corresponds to minimizing

$$\frac{1}{n}\sum_{i=1}^n |\frac{\Delta_i Y_i}{\pi_0} + (1 - \frac{\Delta_i}{\pi_0})(\beta_0 + \beta_1 + \beta_2^T W_i) - \frac{(1-\Delta_i)Y_i}{1-\pi_0} - (1 - \frac{1-\Delta_i}{1-\pi_0})(\beta_0 + \beta_2^T W_i)|^2$$

over $(\beta_0, \beta_1, \beta_2)$, but this is clearly just a modified linear least squares problem. The model matrix is in this case singular, but we discuss in Section 4.5 why any solution will suffice. If the appendix conditions can be verified, which we fully expect, then our resulting locally efficient $\mu_{n,Q_n} = \frac{1}{n}\sum_{i=1}^n D(O_i|\pi_0, Q_n)$ asymptotically makes the best possible use of the working function class $\mathcal{Q}_0$ among all $\mu_{n,Q}$.

A particular consequence is that the empirical efficiency maximization estimate will asymptotically dominate the optimal unadjusted μ_n in (2). The rationale is that from (8), the efficient unadjusted estimator's influence curve is

$$IC_0(O|P_0) = \frac{\Delta Y}{\pi_0} + (1 - \frac{\Delta}{\pi_0})E_{F_0}[Y_1] - \frac{(1-\Delta_i)Y_i}{1-\pi_0} - (1 - \frac{1-\Delta_i}{1-\pi_0})E_{F_0}[Y_0] - \mu(F_0).$$

The function class $\mathcal{Q}_0$ clearly includes this $Q(w,\delta) = \delta E_{F_0}[Y_1] + (1-\delta)E_{F_0}[Y_0]$ for appropriate coefficients. Likewise, the linear least squares fit Q_n will converge to some $Q \in \mathcal{Q}_0$, corresponding to the coefficients minimizing expected squared error $E_{P_0}|Y - Q(W,\Delta)|^2$. By making the most efficient fit of $\mathcal{Q}_0$ for the parameter of interest, we can be more efficient than when using the two special elements associated with unadjusted analysis and linear least squares.

Tsiatis et al. (2000, revised 2006) suggests locally efficient estimation for this treatment effect problem, but proposes decoupling estimation of $Q(F_0)(w,\delta)$ into estimation of outcome regressions $Q_1(F_0)(w) = E_{F_0}[Y|W=w,\Delta=1]$ and $Q_0(F_0)(w) = E_{F_0}[Y|W=w,\Delta=0]$ in the treatment and control group. Fitting a linear least squares in each group reduces to adding a ΔW interaction term to our previous linear model, and it can be shown such a procedure is no less asymptotically efficient than unadjusted estimator (2). When working models $\mathcal{Q}_{1,0}$ and $\mathcal{Q}_{0,0}$ are made for the two regressions $Q_1(F_0)$ and $Q_0(F_0)$, we can of course rejoin to form the index set

$$\mathcal{Q}_0 = \{Q(w,\delta) = \delta Q_1(w) + (1-\delta)Q_0(w) : Q_1 \in \mathcal{Q}_{1,0}, Q_0 \in \mathcal{Q}_{0,0}\}.$$

Empirical efficiency maximization using the rejoined $\mathcal{Q}_0$ would then correspond to targeting the most efficient pair (Q_1, Q_0) in $\mathcal{Q}_{1,0} \times \mathcal{Q}_{0,0}$ for the parameter of interest.

4.2 Excess Risk

With binary outcome $Y \in \{0,1\}$, parameter $\mu(F_0) = P_{F_0}(Y_1 = 1) - P_{F_0}(Y_0 = 1)$ is termed the excess risk. The working index set $\mathcal{Q}_0$ would then more likely be induced by a logistic regression model, and would take the form

$$\mathcal{Q}_0 = \{Q(w,\delta) = \frac{1}{1+\exp(-\beta_0 - \beta_1\delta - \beta_2^T w)} : \beta_0, \beta_1 \in \mathbb{R}, \beta_2 \in \mathbb{R}^p\}. \qquad (10)$$

Standard locally efficient estimation would operate by fitting coefficients with the usual maximum likelihood estimates to form $Q_n \in \mathcal{Q}_0$, and then applying $\mu_{n,Q_n} = \frac{1}{n}\sum_{i=1}^n D(O_i|\pi_0, Q_n)$. Moore and van der Laan (2007) suggest such logistic regression for the excess risk problem, as do Tsiatis et al (2006) (after decoupling). While maximum likelihood would converge to the true $Q(F_0)$ in a correct model, the limiting $Q \in \mathcal{Q}_0$ could lead to a suboptimal parameter estimate under misspecification.

The efficient influence curve clearly suggests we should try to fit $Q(F_0)(w,\delta) = P_{F_0}(Y = 1|W = w, \Delta = \delta)$ as closely as possible, but it is important to define "closely." Should binary regressors minimize Kullback-Leibler divergence, mean squared error, or something else? And does the optimal element in a working model $\mathcal{Q}_{0,1}$ for $Q_1(F_0)(w) = P_{F_0}(Y = 1|W = w, \Delta = 1)$ depend on the element being used in working model

$\mathcal{Q}_{0,0}$, and vice versa? Empirical efficiency maximization answers these questions, by defining the distance from $Q \in \mathcal{Q}_0$ to $Q(F_0)$ through the magnitude of $\sigma^2(Q) = E_{P_0}[D^2(O|\pi_0, Q)] - \mu(F_0)$, the asymptotic variance of the resulting parameter estimate. As we will see shortly, fitting logistic regression coefficients as in (9) would involve not maximizing likelihood, but solving a nonlinear least squares problem.

4.3 Treatment and Control Disease Probabilities

Note that for estimating disease probabilities $\mu_1(F_0) = P_{F_0}(Y_1 = 1) = P_0(Y = 1|\Delta = 1)$ and $\mu_0(F_0) = P_{F_0}(Y_0 = 1) = P_0(Y = 1|\Delta = 0)$ in the treatment and control groups, the problems reduce to two-phase prevalence estimation with constant $P_{G_0}(\Delta = 1|W) = \pi_0(W) = \pi_0$. By another algebraic anomaly, fitting logistic regression logit $P(Y = 1|W, \Delta) = \beta_0 + \beta_1\Delta + \beta_2^T W$ with maximum likelihood and using the locally efficient estimate of $\mu_j(F_0)$ is equivalent to using the plug-in estimate $\mu_{j,n} = \sum_{i=1}^n Q_n(W_i, j)$. Hence, plugging in happens to give consistency and asymptotic linearity. We wouldn't be so fortuitious in general two-phase prevalence estimation, and it is generally important to distinguish locally efficient estimation based on fitting $F_n \in \mathcal{F}_0$ from plugging in $\mu_n = \mu(F_n)$.

4.4 Relative Risks and Odds Ratios

With binary outcomes, the excess risk is not the only way to assess a treatment effect. Consider the (log) relative risk and (log) odds ratio

$$\begin{aligned}
f_{RR}(\mu_1, \mu_0) &= \frac{\mu_1}{\mu_0} = \frac{P_{F_0}(Y_1 = 1)}{P_{F_0}(Y_0 = 1)} \\
f_{log(RR)}(\mu_1, \mu_0) &= \log f_{RR}(\mu_1, \mu_0) \\
f_{OR}(\mu_1, \mu_0) &= \frac{\mu_1}{1-\mu_1} / \frac{\mu_0}{1-\mu_0} = \frac{P_{F_0}(Y_1 = 1)}{P_{F_0}(Y_1 = 0)} / \frac{P_{F_0}(Y_0 = 1)}{P_{F_0}(Y_0 = 0)} \\
f_{log(OR)}(\mu_1, \mu_0) &= \log f_{OR}(\mu_1, \mu_0). \qquad (11)
\end{aligned}$$

To estimate treatment effect $f_{log(OR)}(\mu_1, \mu_0)$, some are tempted to fit the logistic regression model

$$\text{logit } P(Y = 1|W, \Delta) = \beta_0 + \beta_1\Delta + \beta_2^T W,$$

with maximum likelihood, and then report $\beta_{1,n}$. Freedman (2007b) notes the estimator can be inconsistent when the logistic regression model fails. Even when the model holds, Robinson and Jewell (1991) prove that $\beta_{1,n}$ can only lose precision relative to the proper unadjusted estimate, and it appears there is bewilderment about how covariate adjustment should proceed, if at all.

The four treatment effect parameters listed in (11) are not full data population means $E_{F_0}[\psi(X)]$, so we are no longer in the setting of Section 2. Section 6 discusses extensions to parameters solving estimating equations, as in the books of van der Laan and Robins (2003) and Tsiatis (2006), but we are also beyond this scope, because the full data efficient influence curve has no variation independent parametrization in terms of the parameter of interest and nuisance parameters. Handling such difficulties was

one motivation behind our targeted maximum likelihood algorithm in van der Laan and Rubin (2006), later used by Moore and van der Laan (2007) to form covariate-adjusted relative risk and odds ratio estimates. We here consider a different approach.

The efficient influence curves for parameters $\mu_1(F_0)$ and $\mu_0(F_0)$ are

$$\begin{aligned} IC_1(O|\pi_0, Q(F_0), \mu_1(F_0)) &= D_1(O|\pi_0, Q(F_0)) - \mu_1(F_0) \\ &= \frac{\Delta Y}{\pi_0} + (1 - \frac{\Delta}{\pi_0})Q(F_0)(W, 1) - \mu_1(F_0) \end{aligned}$$

and

$$\begin{aligned} IC_0(O|\pi_0, Q(F_0), \mu_0(F_0)) &= D_0(O|\pi_0, Q(F_0)) - \mu_0(F_0) \\ &= \frac{(1-\Delta)Y}{1-\pi_0} + (1 - \frac{1-\Delta}{1-\pi_0})Q(F_0)(W, 0) - \mu_0(F_0). \end{aligned}$$

We consider estimators $\mu_{1,n} = \frac{1}{n}\sum_{i=1}^n D_1(O_i|\pi_0, Q)$ and $\mu_{0,n} = \frac{1}{n}\sum_{i=1}^n D(O_i|\pi_0, Q)$, and then estimate parameter $f(\mu_1, \mu_0)$ with the substitution $f_n = f(\mu_{1,n}, \mu_{0,n})$. Among such substitutions, we then attempt to find the optimal element in working index set $\mathcal{Q}_0$ for our parameter of interest, and estimate $f(\mu_1, \mu_0)$ accordingly. Our limiting element of $\mathcal{Q}_0$ will lead to equal or superior asymptotic variance than when using the unadjusted substitution, or substituting (locally efficient/plug-in) maximum likelihood based logistic regression fits of $\mu_1(F_0)$ and $\mu_0(F_0)$.

For the previously listed treatment effects, the delta method can trivially be used to derive influence curves for substitution estimators $f_n = f(\mu_{1,n}, \mu_{0,n})$ of $f(\mu_1, \mu_0)$, which are

$$IC_{RR}(O) = IC_1(O|\pi_0, Q, \mu_1)\frac{\partial f_{RR}}{\partial \mu_1}(\mu_1, \mu_0) + IC_0(O|\pi_0, Q, \mu_0)\frac{\partial f_{RR}}{\partial \mu_0}(\mu_1, \mu_0)$$

$$IC_{log(RR)}(O) = IC_1(O|\pi_0, Q, \mu_1)\frac{\partial f_{log(RR)}}{\partial \mu_1}(\mu_1, \mu_0) + IC_0(O|\pi_0, Q, \mu_0)\frac{\partial f_{log(RR)}}{\partial \mu_0}(\mu_1, \mu_0)$$

$$IC_{OR}(O) = IC_1(O|\pi_0, Q, \mu_1)\frac{\partial f_{OR}}{\partial \mu_1}(\mu_1, \mu_0) + IC_0(O|\pi_0, Q, \mu_0)\frac{\partial f_{OR}}{\partial \mu_0}(\mu_1, \mu_0)$$

$$IC_{log(OR)}(O) = IC_1(O|\pi_0, Q, \mu_1)\frac{\partial f_{log(OR)}}{\partial \mu_1}(\mu_1, \mu_0) + IC_0(O|\pi_0, Q, \mu_0)\frac{\partial f_{log(OR)}}{\partial \mu_0}(\mu_1, \mu_0),$$

where the partial derivatives are given by

$$[\frac{\partial f_{RR}}{\partial \mu_1}, \frac{\partial f_{RR}}{\partial \mu_0}](\mu_1, \mu_0) = [\mu_0^{-1}, -\mu_1\mu_0^{-2}]$$

$$[\frac{\partial f_{log(RR)}}{\partial \mu_1}, \frac{\partial f_{log(RR)}}{\partial \mu_0}](\mu_1, \mu_0) = [\mu_1^{-1}, -\mu_0^{-1}]$$

$$[\frac{\partial f_{OR}}{\partial \mu_1}, \frac{\partial f_{OR}}{\partial \mu_0}](\mu_1, \mu_0) = [\frac{(1-\mu_0)\{(1-\mu_1)+\mu_1\}}{(1-\mu_1)^2\mu_0}, \frac{\mu_1\{-(1-\mu_0)-\mu_0\}}{(1-\mu_1)\mu_0^2}]$$

$$[\frac{\partial f_{log(OR)}}{\partial \mu_1}, \frac{\partial f_{log(OR)}}{\partial \mu_0}](\mu_1, \mu_0) = [\mu_1^{-1} + (1-\mu_1)^{-1}, -\mu_0^{-1} - (1-\mu_0)^{-1}].$$

The asymptotic variance of $f(\mu_{1,n}, \mu_{0,n})$ is thus $\sigma^2(Q) = E_{P_0}[IC^2(O|\pi_0, Q, \mu_1, \mu_0)]$. Based on preliminary estimates $\hat{\mu}_{0,n}$ and $\hat{\mu}_{1,n}$, this could be approximated empirically

with $\frac{1}{n}\sum_{i=1}^{n} IC^2(O_i|\pi_0, Q, \hat{\mu}_{0,n}, \hat{\mu}_{1,n})$, and minimized over working index set $\mathcal{Q}_0$. Empirical efficiency maximization is thus to select

$$Q_n = \text{argmin}_{Q\in\mathcal{Q}_0} \frac{1}{n}\sum_{i=1}^{n} IC^2(O_i|\pi_0, Q, \hat{\mu}_{1,n}, \hat{\mu}_{0,n}),$$

and then estimate the parameter of interest with

$$f_n = f(\frac{1}{n}\sum_{i=1}^{n} D_1(O_i|\pi_0, Q_n), \frac{1}{n}\sum_{i=1}^{n} D_0(O_i|\pi_0, Q_n)).$$

The preliminary $\hat{\mu}_{1,n}$ and $\hat{\mu}_{0,n}$ will generally only need to be consistent for f_n to be asymptotically equivalent to the substitution estimator using optimal $Q \in \mathcal{Q}_0$. Unadjusted preliminary fits would suffice for this purpose. Section 6 provides more details on the extension of empirical efficiency maximization to such substitution estimators.

4.5 Implementation

For the parameters we've considered estimating in randomized experiments, our optimization problems can be carried out with nonlinear least squares. Suppose our working model $\mathcal{Q}_0$ for the outcome regression $Q(F_0)(w, \delta) = E_{F_0}[Y|W = w, \Delta = w]$ is parametrized by a finite dimensional β, as when using linear or logistic regression models. Using surrogate responses $\{Y_i^\star = Y^\star(O_i)\}_{i=1}^n$ and a function $h_\beta(O)$ of the observed data $O = (W, \Delta, Y)$, empirical efficiency maximization reduces to finding β by minimizing $\frac{1}{n}\sum_{i=1}^{n} |Y_i^\star - h_\beta(O_i)|^2$. The surrogate responses $Y^\star(O)$ are given by

Mean treatment response: $\frac{\Delta Y}{\pi_0}$

Mean control response: $\frac{(1-\Delta)Y}{1-\pi_0}$

Treatment effect/excess risk: $\frac{\Delta Y}{\pi_0} - \frac{(1-\Delta)Y}{1-\pi_0}$

Relative risk: $(\frac{\Delta Y}{\pi_0} - \hat{\mu}_{1,n})\frac{\partial f_{RR}}{\partial \mu_1}(\hat{\mu}_{1,n}, \hat{\mu}_{0,n}) + (\frac{(1-\Delta)Y}{1-\pi_0} - \hat{\mu}_{0,n})\frac{\partial f_{RR}}{\partial \mu_0}(\hat{\mu}_{1,n}, \hat{\mu}_{0,n})$

Log(RR): $(\frac{\Delta Y}{\pi_0} - \hat{\mu}_{1,n})\frac{\partial f_{log(RR)}}{\partial \mu_1}(\hat{\mu}_{1,n}, \hat{\mu}_{0,n}) + (\frac{(1-\Delta)Y}{1-\pi_0} - \hat{\mu}_{0,n})\frac{\partial f_{log(RR)}}{\partial \mu_0}(\hat{\mu}_{1,n}, \hat{\mu}_{0,n})$

Odds ratio: $(\frac{\Delta Y}{\pi_0} - \hat{\mu}_{1,n})\frac{\partial f_{OR}}{\partial \mu_1}(\hat{\mu}_{1,n}, \hat{\mu}_{0,n}) + (\frac{(1-\Delta)Y}{1-\pi_0} - \hat{\mu}_{0,n})\frac{\partial f_{OR}}{\partial \mu_0}(\hat{\mu}_{1,n}, \hat{\mu}_{0,n})$

Log(OR): $(\frac{\Delta Y}{\pi_0} - \hat{\mu}_{1,n})\frac{\partial f_{log(OR)}}{\partial \mu_1}(\hat{\mu}_{1,n}, \hat{\mu}_{0,n}) + (\frac{(1-\Delta)Y}{1-\pi_0} - \hat{\mu}_{0,n})\frac{\partial f_{log(OR)}}{\partial \mu_0}(\hat{\mu}_{1,n}, \hat{\mu}_{0,n}),$

and the fitted values $h_\beta(O)$ are given by

$$\begin{aligned}
&\text{Mean treatment response: } -(1-\frac{\Delta}{\pi_0})Q_\beta(W,1)\\
&\text{Mean control response: } -(1-\frac{1-\Delta}{1-\pi_0})Q_\beta(W,0)\\
&\text{Treatment effect/excess risk: } -(1-\frac{\Delta}{\pi_0})Q_\beta(W,1)+(1-\frac{1-\Delta}{1-\pi_0})Q_\beta(W,0)\\
&\text{RR: } -(1-\frac{\Delta}{\pi_0})\frac{\partial f_{RR}}{\partial \mu_1}(\hat\mu_{1,n},\hat\mu_{0,n})Q_\beta(W,1)-(1-\frac{1-\Delta}{1-\pi_0})\frac{\partial f_{RR}}{\partial \mu_0}(\hat\mu_{1,n},\hat\mu_{0,n})Q_\beta(W,0)\\
&\text{Log(RR): } -(1-\frac{\Delta}{\pi_0})\frac{\partial f_{log(RR)}}{\partial \mu_1}(\hat\mu_{1,n},\hat\mu_{0,n})Q_\beta(W,1)\\
&\qquad -(1-\frac{1-\Delta}{1-\pi_0})\frac{\partial f_{log(RR)}}{\partial \mu_0}(\hat\mu_{1,n},\hat\mu_{0,n})Q_\beta(W,0)\\
&\text{OR: } -(1-\frac{\Delta}{\pi_0})\frac{\partial f_{OR}}{\partial \mu_1}(\hat\mu_{1,n},\hat\mu_{0,n})Q_\beta(W,1)-(1-\frac{1-\Delta}{1-\pi_0})\frac{\partial f_{OR}}{\partial \mu_0}(\hat\mu_{1,n},\hat\mu_{0,n})Q_\beta(W,0)\\
&\text{Log(OR): } -(1-\frac{\Delta}{\pi_0})\frac{\partial f_{log(OR)}}{\partial \mu_1}(\hat\mu_{1,n},\hat\mu_{0,n})Q_\beta(W,1)\\
&\qquad -(1-\frac{1-\Delta}{1-\pi_0})\frac{\partial f_{log(OR)}}{\partial \mu_0}(\hat\mu_{1,n},\hat\mu_{0,n})Q_\beta(W,0).
\end{aligned}$$

Note that some nonlinear least squares software will give error messages for such minimizations, because of a singular gradient of the objective function. This should not be of great concern. The problem is that different values of β or Q_β can lead to equivalent influence curves $IC(O|\pi_0,Q_\beta,\mu(F_0))=D(O|\pi_0,Q_\beta)-\mu(F_0)$. Our estimator's asymptotics are determined by the limiting value of $D(\cdot|\pi_0,Q_{\beta_n})$, and how the functions $\mathcal{D}_0=\{D(\cdot|\pi_0,Q):Q\in\mathcal{Q}_0\}$ are parametrized is irrelevant.

One may find that their β_n doesn't converge, or fails to converge to the true β_0 after generating data according to a working model. By themselves such phenomena aren't necessarily ominous, and in such cases the first step in checking convergence of the function $D(\cdot|\pi_0,Q_n)$. For the (log) relative risk and (log) odds ratio, the same disclaimer of course applies even though influence curves aren't of the form $IC(O|\pi_0,Q,\mu(F_0))=D(O|\pi_0,Q)-\mu(F_0)$, and we must examine the behavior of $IC(\cdot|\pi_0,Q_{\beta_n},\hat\mu_{1,n},\hat\mu_{0,n})$.

4.6 Asymptotic Variance Calculations

We again compared estimators by examining asymptotic variance, after simulating an experiment as follows:

$$\begin{aligned}
&W\sim N(0,1)\\
&P_{G_0}(\Delta=1|W,Y)=P_{G_0}(\Delta=1)=\pi_0\\
&P_0(Y=1|W,\Delta)=\frac{1}{1+\exp(-\Delta W+(1+\Delta)W)}.
\end{aligned}$$

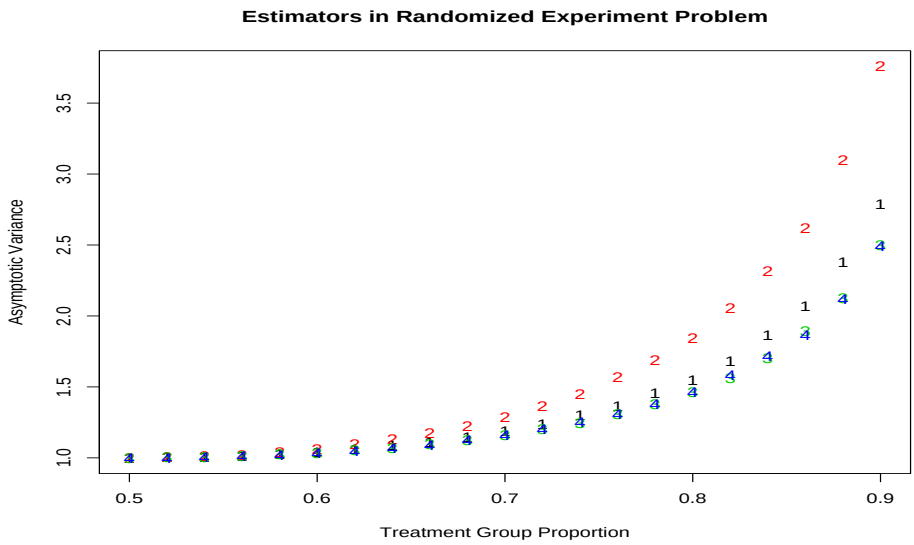

Figure 2: $1-4$ represent the efficient unadjusted estimator (2) of excess risk, a locally efficient estimator fitting a misspecified logistic regression model's coefficients with maximum likelihood, the empirical efficiency estimator, and the fully efficient estimator. Performance is similar. But as the treatment probability increases, standard locally efficient covariate adjustment becomes worse than no adjustment, while empirical efficiency maximization essentially leads to full efficiency.

We varied the treatment assignment probability from $\pi_0 = \frac{1}{2}$ to $\pi_0 = \frac{9}{10}$ in steps of 0.02. Here covariate W was positively associated with outcome Y in the treatment group ($\Delta = 1$), while negatively associated in the control group ($\Delta = 0$). We considered estimating the excess risk $\mu(F_0) = P_0(Y = 1|\Delta = 1) - P_0(Y = 1|\Delta = 0)$, which in this case was zero for any value of π_0.

Based on a sample of size $n = 100,000$ we computed the limiting influence curves of locally efficient estimators based on fitting a misspecified logistic regression model with maximum likelihood and empirical efficiency maximization. For the latter, we minimized $\frac{1}{n}\sum_{i=1}^{n} D^2(O_i|\pi_0, Q_{(\gamma,\alpha,\beta)})$ with the **nlminb()** in the R language where

$$Q_{(\alpha,\gamma,\beta)}(w,\delta) = \frac{1}{1+\exp(-\gamma-\alpha\delta-\beta w)}.$$

By design, we also knew the influence curves of the unadjusted estimator (2), and the efficient estimator. Based on a new sample of size $n = 100,000$, we computed asymptotic variances by finding the empirical means of the squared influence curves.

Figure 2 displays results. There didn't appear to be large differences between estimators, but the structure is worth noting. Near $\pi_0 = \frac{1}{2}$, performance appeared identical. But as the treatment assignment probability increased, unadjusted analysis appeared superior than standard locally efficient analysis, while empirical efficiency maximization was indistinguishable from efficient estimation.

5 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, covariates are 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 $X = (W, T) \sim F_0$, which we might not directly observe because some subjects cannot be monitored until failure. We will let the coarsening mechanism $G_0(\cdot)$ denote the distribution function of censoring variable C, and $\bar{G}(c) = 1 - G_0(c) = P(C > c)$ the censoring variable's survival curve, and interchangeably refer to this function as the censoring mechanism.

5.1 Marginal Survival

Suppose interest lies in estimating the full data parameter $\mu(F_0) = P_{F_0}(T > t)$, the survival probability at a single time t, such as five-year survival. As censoring is often caused by study termination, we also assume it to be completely independent of a subject's covariates and failure time. This is written

$$X = \{W, T\} \perp C, \tag{12}$$

and implies coarsening at random.

In fact, this independence assumption (12) is implicitly made by those who ignore clinically informative covariates and fit the survival curve with the well-known estimator of Kaplan and Meier (1958), which is what would most likely be done in practice. Actually, Kaplan-Meier requires the weaker marginal independence $\{T \perp C\}$, but for covariates predictive of failure time, it is hard to imagine how this could be clinically justified without the stronger complete independence.

If a component of W is associated with censoring but not survival, it can be discarded to satisfy (12), and our estimator will suffer no precision loss. We also assume that $\bar{G}_0(\tilde{T} \vee t)$ is bounded away from zero with probability one, and survival can be truncated shortly after t to ensure this without altering the parameter of interest.

While convenient, ignoring informative covariates can lead to a serious loss in efficiency, and anyone who has viewed Kaplan-Meier confidence bands can attest that the estimator's precision often leaves much to be desired. The rationale is that clinically predictive measurements provide extra information about failure times of individuals lost to censoring. Empirical efficiency maximization is thus something to consider when Kaplan-Meier is an appropriate temptation.

The efficient influence curve for survival probability $\mu(F_0) = P_{F_0}(T > t)$ is

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

for the doubly robust mapping of $\psi(X) = I(T > t)$ given by

$$D(O|G_0, Q(F_0)) = \frac{\Delta I(\tilde{T} > t)}{\bar{G}_0(\tilde{T}_-)} + \int \frac{Q(F_0)(c, W)}{\bar{G}_0(c)} dM(c),$$

where

$$Q(F) : (W, c) \to P_F(T > t | T > c, W)$$

maps full data distributions $F \in \mathcal{F}$ to functions from $\mathcal{W} \times \mathcal{C}$ to $[0, 1]$. Here

$$\begin{aligned} M(c) &= N(c) - A(c) \\ N(c) &= I(\tilde{T} \leq c, \Delta = 0) \\ A(c) &= \int_{-\infty}^{c} I(\tilde{T} \geq u) \frac{dG(u)}{\bar{G}(u_-)}. \end{aligned}$$

The martingale $M(\cdot)$ is built from the Doob-Meyer decomposition of the counting process $N(\cdot)$ jumping at an observed censoring time, and $A(\cdot)$ 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).

Once again, the robustness result (5) implies that $\mu_{n,Q} = \frac{1}{n} \sum_{i=1}^{n} D(O_i|G_0, Q)$ would be unbiased for any $Q \in \mathcal{Q} = \{Q(F) : F \in \mathcal{F}\}$. 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. Unlike in two-phase prevalence studies or randomized experiments, the unbiased $\mu_{n,Q}$ couldn't be applied directly, because the coarsening mechanism G_0 is not exactly known. It is now a nuisance parameter, and we consider efficiently estimating it with G_n based on the Kaplan-Meier fit. As discussed in Section 2, efficiently estimating the coarsening mechanism can never hurt asymptotically, and $\frac{1}{n} \sum_{i=1}^{n} D(O_i|G_n, Q)$ will be no less asymptotically efficient than the (unavailable) unbiased $\frac{1}{n} \sum_{i=1}^{n} D(O_i|G_0, Q)$.

We could attempt constructing a fully efficient estimator for this problem, by consistently estimating $Q(F_0)(c, W) = P_{F_0}(T > t | T > c, W)$ with a Q_n, and then applying $\mu_n = \frac{1}{n} \sum_{i=1}^{n} D(O_i|G_n, Q_n)$. While the estimator's asymptotic variance could be much smaller than Kaplan-Meier's, asymptotic efficiency isn't always advisable. Consistently fitting $Q(F_0)$ could be a difficult smoothing exercise, and Q_n might not well approximate $Q(F_0)$ for reasonable sample sizes.

Chapter 3 of van der Laan and Robins (2003) reviews methods for locally efficient estimation in general right censored data structures, which can be applied when estimating a marginal survival curve. Along these lines, Zeng (2004) considers locally efficient estimation of marginal survival based on a combination of Cox modeling and smoothing. Similar semiparametric working models are proposed in Satten et al. (2001) and Scharfstein and Robins (2002) for marginal survival. It is fair to say that work in this area has primarily been directed toward using covariates to correct for dependent censoring $\neg\{T \perp C\}$, rather than to increase efficiency. In fact, Satten et al. state that when clinically informative covariates are believed to have no effect on censoring as in (12), their estimator reduces to the Kaplan-Meier curve.

In locally efficient estimation, one fits a relatively small working model $\mathcal{F}_0$ with F_n for the full data distribution, and takes $Q_n(c, W) = P_{F_n}(T > t | T > c, W)$ for use in

$\mu_n = \frac{1}{n}\sum_{i=1}^n D(O_i|G_n, Q_n)$. Because only the conditional law $\mathcal{L}(T|W)$ matters for the $Q(F_0)$ appearing in the efficient influence curve, it suffices to model this conditional distribution, and ignore the marginal covariate distribution. Under our independence assumption (12), asymptotic efficiency is achieved for a correctly specified working model $\mathcal{F}_0$, while misspecification does not compromise asymptotic linearity.

In almost all survival analysis applications, the most frequently used working model for conditional survival time is Cox's (1972) proportional hazards model. Parametrized by a coefficient vector β and baseline cumulative hazard function $\Lambda_0(\cdot)$, it induces the working index set

$$\mathcal{Q}_0 = \{Q_{\beta,\Lambda_0}(c, W) = \exp(-e^{\beta^T W}\{\Lambda_0(t \vee c) - \Lambda_0(c)\}) : \beta \in \mathbb{R}^p, \Lambda_0(\cdot)\}. \tag{13}$$

The coefficient vector β is typically fit through Cox's well-known partial likelihood technique, while the baseline cumulative hazard $\Lambda_0(\cdot)$ is fit using Breslow's (1974) estimator. As in previous examples, we stress that locally efficient estimation based on the Cox model fit of $Q(F_0)$ is entirely different from estimating marginal survival with the Cox model plug-in estimator.

While partial likelihood converges to the true model parameters if the proportional hazards model is correct, there is no reason to believe such estimation will generally converge to the optimal $\mathcal{Q}_0$ for estimating marginal survival. We represent asymptotic variance $\sigma^2(Q)$ in the following theorem, proven in the appendix.

Theorem 2. *Let $\phi(c, \tilde{T}) = 1 - A(c) + \lim_{u\uparrow c} A(u)$, where our notation expresses that the compensator $A(\cdot)$ is a random process depending on $\tilde{T}$. The limit in the definition will exist as the compensator is right continuous. Then,*

$$\begin{aligned}\sigma^2(Q) &= Var_{P_0}(D(O|G_0, Q)) = E_{P_0}[D^2(O|G_0, Q)] - \mu^2(F_0) \\ &= C(P_0) + E_{(O,C)\sim P_0\times G_0}[\alpha(\tilde{T}, C)Q^2(C, W) - 2\beta(\Delta, \tilde{T}, C)Q(C, W)],\end{aligned}$$

for

$$C(P_0) = -\mu^2(F_0) + E_{P_0}|\frac{\Delta I(\tilde{T} > t)}{\bar{G}_0(\tilde{T}_-)}|^2$$

not depending on Q, and

$$\alpha(\tilde{T}, C) = \frac{I(\tilde{T} \geq C)\phi(C, \tilde{T})}{\bar{G}_0^2(C)\bar{G}_0(C_-)}$$
$$\beta(\Delta, \tilde{T}, C) = \frac{\Delta I(\tilde{T} > t)I(\tilde{T} \geq C)}{\bar{G}_0(\tilde{T}_-)\bar{G}_0(C)\bar{G}_0(C_-)}.$$

We can take $\phi(C, \tilde{T}) = 1$ if the censoring variable C has a density under G_0. In this case, completing the square gives the immediate implication that $\sigma^2(Q)$ is monotone increasing in the weighted mean squared error,

$$E_{(O,C)\sim P_0\times G_0}\left\{\frac{I(\tilde{T} \geq C)}{\bar{G}_0^3(C)}|\frac{\Delta I(\tilde{T} > t)\bar{G}_0(C)}{\bar{G}_0(\tilde{T})} - Q(C, W)|^2\right\}.$$

The theorem reveals that for a continuous censoring variable C, the optimal $Q \in \mathcal{Q}_0$ minimizes a weighted mean squared error. The empirical efficiency maximization idea is to select $Q \in \mathcal{Q}_0$ by minimizing $\frac{1}{n}\sum_{i=1}^{n} D^2(O_i|G_n, Q)$. We can equivalently seek the optimal working model element by replacing the unknown data generating distribution P_0 with the empirical distribution $\mathbb{P}_n$ in $E_{(O,C)\sim P_0\times G_0}[\alpha(\tilde{T}, C)Q^2(C, W) - 2\beta(\Delta, \tilde{T}, C)Q(C, W)]$, and plugging-in an efficient estimate G_n for the censoring mechanism G_0. The objective function then becomes a double integral with respect to $\mathbb{P}_n$ and G_n, which can easily be evaluated by Monte Carlo. As in two-phase prevalence estimation, we take the liberty of claiming this modified optimization falls under the rubric of empirical efficiency maximization. The procedure 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 $\{c_b\}_{b=1}^{B}$ from G_n.
3. Form $\alpha_b = \alpha(\tilde{t}_b, c_b)$ and $\beta_b = \beta(\delta_b, \tilde{t}_b, c_b)$ for $b = 1, ..., B$, with G_n substituted for G_0.
4. Choose $Q \in \mathcal{Q}_0$ to minimize $\frac{1}{B}\sum_{b=1}^{B}\{\alpha_b Q^2(c_b, w_b) - 2\beta_b Q(c_b, w_b)\}$.

When we know the censoring time C has a density, our theorem tells us we can reduce the algorithm to:

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 $\{c_b\}_{b=1}^{B}$ from G_n.
3. Form weights $a_b = \frac{I(\tilde{t}_b \geq c_b)}{G_n^3(c_b)}$ and surrogate responses $y_b = \frac{\delta_b I(\tilde{t}_b > t)\bar{G}_n(c_b)}{\bar{G}_n(\tilde{t}_b)}$ for $b = 1, ..., B$.
4. Choose $Q \in \mathcal{Q}_0$ to minimize the weighted squared error $\frac{1}{B}\sum_{b=1}^{B} a_b|y_b - Q(c_b, w_b)|^2$.

When the working model $\mathcal{Q}_0$ is induced by a Cox model as in (13), it may appear that empirical efficiency maximization necessitates solving an infinite dimensional minimization problem, as we have to minimize our objective function over all monotone baseline hazards $\Lambda_0(\cdot)$. Fortunately, inspection of our algorithm reveals we only need to evaluate $\Lambda_0(\cdot)$ at survival endpoint t, and the unique censoring times in $\{c_b\}_{b=1}^{B}$ occuring before time t. We thus only have to find a finite number of nonnegative hazards, and the problem reduces to finite dimensional weighted nonlinear least squares, with nonnegativity constraints for some parameters.

Interestingly, this Cox model fit will be quite different from the usual fit. Breslow's baseline cumulative hazard estimator only places mass at the unique failure times, while our proposal is to place mass only at t and times $c \leq t$ in the support of estimated censoring mechanism G_n. To intuitively understand why this is so, we need to observe that $Q(c, W)$ will only be evaluated in $D(O|G_0, Q)$ at times $c \leq t$ in the support of the censoring variable C. Estimated $P_0(T > c|W)$ will not affect the locally efficient estimator if c is greater than t or is outside the censoring variable's support. However, fits at such irrelevant times surely contribute to the (partial) likelihood, and standard

methods can be suboptimal as a result. While not necessarily leading to a desirable fit of the conditional distribution $\mathcal{L}(T|W)$, our procedure is targeted toward estimating the marginal survival parameter of interest. However, we must caution the appendix templates have not yet been used to derive our empirical efficiency maximization estimator's asymptotics for the working Cox model.

By examining the efficient influence curve's $Q(F_0)(c, W) = P_{F_0}(T > t|T > c, W)$, and recalling the Kaplan-Meier estimator is efficient when no covariates are measured, it immediately follows that the Kaplan-Meier estimator is asymptotically equivalent to $\mu_{n,Q} = \frac{1}{n}\sum_{i=1}^{n} D(O_i|G_0, Q)$ with $Q(c, W) = Q(c) = P_{F_0}(T > t|T > c)$. The working index set $\mathcal{Q}_0$ induced by the Cox model clearly includes this function at $\beta = 0$ and appropriate baseline cumulative hazard function $\Lambda_0(\cdot)$. Consequently, the efficient choice of $Q \in \mathcal{Q}_0$ will at least lead to Kaplan-Meier efficiency. For general full data working models $\mathcal{F}_0$, there is no reason why this must hold for $Q(F) \in \mathcal{Q}_0$ at the $F \in \mathcal{F}_0$ minimizing Kullback-Leibler divergence from F_0.

Empirical efficiency maximization could also be utilized for a covariate-adjusted estimate of mean failure time $E_{F_0}[T]$ (or more likely $E_{F_0}[T \wedge t]$ for identifiability). The extensions in Section 6.4 show how our method could as well be used to approximate quantiles of $\mathcal{L}(T)$, or other functionals. Using a substitution estimate based on a covariate-adjusted survival curve, we could additionally use the technique of Section 6.2 to estimate the cumulative hazard at a point.

A drawback of our new locally efficient method is that we've targeted a univariate parameter. Section 6 discusses multivariate extensions, but it is not clear how we could best use a full data working model to fit an entire survival curve.

5.2 Asymptotic Variance Calculations

We explored estimation techniques by generating data structures as follows:

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

Here η was a model misspecification parameter. 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, and 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. We considered estimating the five-year survival $\mu(F_0) = P_{F_0}(T > 5)$.

In these simple simulations, $D(O|G_0, Q)$ only required evaluating $Q(c, W)$ at $c = 3$, so we reduce notational overhead by writing $Q(W)$. From knowledge of the censoring mechanism, which we here assume, 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, $\sigma^2(Q) = E_{P_0}[D^2(O|G_0, Q)] - \mu^2(F_0)$ is monotone in the weighted squared error $E_{P_0}[I(\tilde{T} \geq 3)|2I(\tilde{T} > 5) - Q(W)|^2]$. Empirical efficiency maximization thus reduced to selecting a working set $\mathcal{Q}_0$, and finding the $Q_n \in \mathcal{Q}_0$ to minimize a weighted least squares. Using a working Cox model for local efficiency, the working index set $\mathcal{Q}_0$ 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 $\beta_0 = \log(\Lambda_0(5) - \Lambda_0(3))$.

Hence, we considered estimators $\mu_{n,(\beta_0,\beta_1)} = \frac{1}{n}\sum_{i=1}^n D(O_i|G_0, Q_{\beta_0,\beta_1})$. For each model misspecification parameter η, we evaluated the limiting (β_0, β_1) of partial likelihood maximization and empirical efficiency maximization, through generating a dataset with sample size $n = 100,000$. As in the two-phase study example, the weighted nonlinear least squares in empirical efficiency maximization was solved with the **nls()** function in the R language. This simulated dataset also allowed us to find survival probability $\mu(F_0) = P_{F_0}(T > 5)$ via Monte Carlo. Using a new independent sample of the same size, we then computed asymptotic variances for the two estimators by evaluating $\frac{1}{n}\sum_{i=1}^n D^2(O_i|G_0, Q) - \mu^2(F_0)$. Additionally, we computed asymptotic variances for the Kaplan-Meier influence curve's $Q(W) = P_{F_0}(T > 5|T > 3)$, and the efficient influence curve's $Q(F_0)(W) = P_{F_0}(T > 5|T > 3, W)$. 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 3. 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 important phenomenon shown in both this simulation and previous simulations is that even with a misspecified working model, elements of the working model can lead to estimators that are extremely close to being fully efficient for our parameter, and likelihood based estimates do not always converge to these elements.

5.3 Comparing Survival Distributions

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

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

In a randomized trial, A could correspond to a treatment indicator. Interest might lie in testing whether A is associated with survival, meaning the conditional law $\mathcal{L}(T|A = 1)$ differs from $\mathcal{L}(T|A = 0)$. As our method can estimate a survival probability at a point $P_{F_0}(T > t)$, the cumulative hazard at a point, mean survival, or

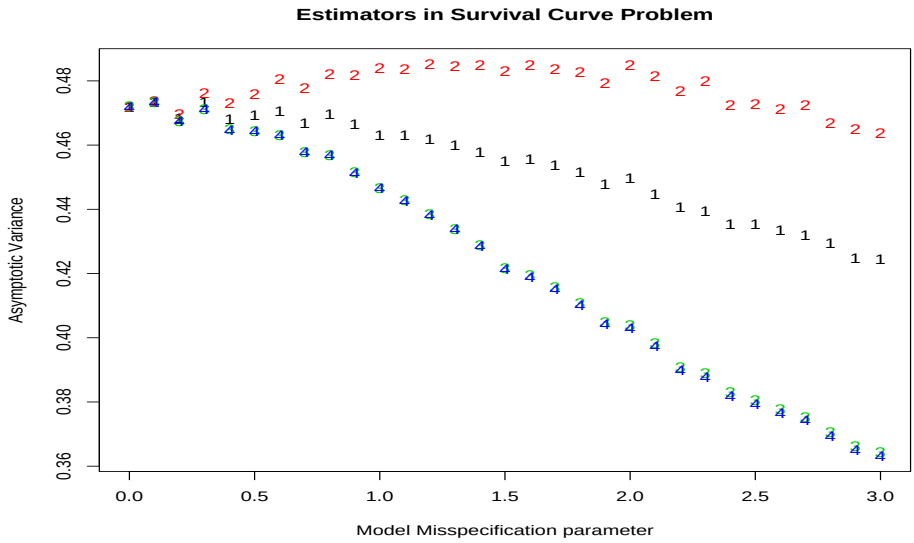

Figure 3: $1-4$ represent the Kaplan-Meier estimator that ignores covariates, a locally efficient estimator based on a misspecified Cox model fit with the usual partial likelihood technique and Breslow baseline cumulative hazard, an empirical efficiency maximization estimator based on the Cox model, and the efficient estimator. The Kaplan-Meier estimator ignoring covariates outperformed the likelihood based locally efficient procedure. Empirical efficiency maximization essentially led to full efficiency.

median survival, it would be straightforward to estimate the difference in such quantities between two distributions, such as $\mu(F_0) = P_{F_0}(T > t|A = 1) - P_{F_0}(T > t|A = 0)$. We could then test the null hypothesis that $\mu(F_0) = 0$.

While the parameters we mentioned would be legitimate for testing, they are not commonly used in survival analysis. Let $\Lambda_1(\cdot)$ and $\Lambda_0(\cdot)$ denote the cumulative hazard functions for the two distributions. Chapter 7 of Fleming and Harrington (1991) reviews how many test statistics are $n^{-1/2}$-scaled versions of

$$\mu_n = \int K_n(t) d\{\Lambda_{1,n}(t) - \Lambda_{0,n}(t)\}, \tag{14}$$

for Nelson-Aalen fits of the baseline hazards. For $K_n \to K(P_0)$, such μ_n can converge to parameter

$$\mu(P_0) = \int K(P_0)(t) d\{\Lambda_1(t) - \Lambda_0(t)\}.$$

The popular logrank statistic is of this form, with

$$K(P_0)(t) = P_0(A = 1)P_0(A = 0)\frac{P_0(\tilde{T} > t|A = 1)P_0(\tilde{T} > t|A = 0)}{P_0(\tilde{T} > t)}.$$

We could attempt to more precisely estimate this parameter with covariate adjustment, and thus increase power when testing $H_0 : \mu(P_0) = 0$. Unfortunately, several complications arise. Because the function $K(P_0)(\cdot)$ can depend on the censoring

mechanism for weighted logrank statistics, parameter $\mu(P_0)$ often won't be a full data parameter $\mu(F_0)$. Even without this problem, we would still be outside the general formulation in Section 2, because the parameter wouldn't be a full data population mean $\mu(F_0) = E_{F_0}[\psi(X)]$. Moreover, the full data efficient influence curve would have no variation independent parametrization in terms of the parameter of interest and a nuisance parameter, so wouldn't be covered by the theory of van der Laan and Robins (2003) or Tsiatis (2006), and we couldn't apply our extended estimating function based empirical efficiency maximization of Section 6.4. However, as with the (log) relative risk and (log) odds ratio in Section 4, we could consider covariate-adjusted substitution estimators, discussed more generally in Section 6.2.

For F in full data working model $\mathcal{F}_0$, the previous subsection considered covariate-adjusted marginal survival estimates at a time t. By using the estimator at different times t, one arrives at a (possibly improper) estimator of the entire curve. This could be carried out for the $\{A = 1\}$ and $\{A = 0\}$ groups. Marginal survival fits $S_{1,n}$ and $S_{0,n}$ could then be mapped into covariate-adjusted baseline hazard fits via $\Lambda_{j,n}(\cdot) = \int_{-\infty}^{\cdot} \frac{d\bar{S}_{j,n}(t)}{S_{j,n}(t_-)}$, which would be entered into the μ_n of (14) for a covariate-adjusted parameter estimate.

If we could derive this estimator's influence curve $IC(O|G_0, F, \eta(P_0))$, we could estimate nuisance parameter $\eta(P_0)$, and empirically approximate asymptotic variance $\sigma^2(F) = E_{P_0}[IC^2(O|G_0, F, \eta(P_0))]$ with $\hat{\sigma}^2(F) = \frac{1}{n}\sum_{i=1}^{n} IC^2(O_i|G_n, F, \eta_n)$, and attempt to minimize over F in a full data working model $\mathcal{F}_0$ to make the optimal covariate adjustment for the parameter of interest $\mu(P_0)$.

At fixed alternatives $\mu(P_0) \neq 0$, most tests have power converging to one, and hence interest has traditionally concerned power at local alternatives. It can be shown that at local alternatives $\Lambda_1(t) = \Lambda_0(t) + n^{-1/2}h(t)$, the asymptotics would be as if $K(P_0)(\cdot)$ were fixed and known, and we can finesse the issue of estimating a parameter depending on the coarsening mechanism.

Obviously, this is a very preliminary preview of what empirical efficiency maximization would entail for comparing survival distributions, and we intend to elaborate in the near future. To mention related work, Lu (2006) gave a talk presenting covariate-adjusted tests asymptotically outperforming the logrank procedure when the $\mathcal{L}(T|A = 1)$ and $\mathcal{L}(T|A = 0)$ distributions obey proportional hazards. The forthcoming article of Lu and Tsiatis (2007) will apparently expound.

6 Extensions

In Section 2, we restricted attention to estimating univariate full data population means of the form $\mu(F_0) = E_{F_0}[\psi(X)] \in \mathbb{R}$. This was meant to be expository, and empirical efficiency maximization in this limited setting can go a long way. Slight modifications are required to attack more general parameters, and in this section we informally sketch how one might proceed.

6.1 Multivariate Parameters

A drawback of empirical efficiency maximization is that there is no straightforward generalization to estimating multivariate parameters $\mu(F_0) = [\mu_1(F_0), ..., \mu_k(F_0)]^T \in \mathbb{R}^k$. It is not necessarily true that within a misspecified full data submodel $\mathcal{F}_0 \subset \mathcal{F}$, there is a single F making $\frac{1}{n}\sum_{i=1}^n [D_1(O_i|G_0, Q(F)), ..., D_k(O_i|G_0, Q(F))]^T$ as efficient as possible, meaning this is the best $F \in \mathcal{F}_0$ for approximating any linear combination of $[\mu_1(F_0), ..., \mu_k(F_0)]^T$.

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 is to represent the influence curve of multivariate parameter estimate $\mu_n = [\mu_{n,1}, ..., \mu_{n,k}]^T$ as the vector

$$IC(O|G_0, Q(F), \mu(F_0)) = [IC_1(O|G_0, Q(F), \mu(F_0)), ..., IC(O|G_0, Q(F), \mu(F_0))]^T.$$

With a consistent preliminary parameter estimate $\hat{\mu}_n$, the asymptotic covariance matrix (of the estimator applied with known coarsening mechanism) could be approximated empirically with

$$\hat{\Sigma}(F) = \frac{1}{n}\sum_{i=1}^n IC(O_i|G_n, Q(F), \hat{\mu}_n) IC(O_i|G_n, Q(F), \hat{\mu}_n)^T.$$

Empirical efficiency maximization could then operate by defining a norm $\|\cdot\|$ on covariance matrices, and trying to minimize $\|\hat{\Sigma}(F)\|$ over a working model $\mathcal{F}_0 \subset \mathcal{F}$. To reiterate, there might be no "best" F for estimating all components of the parameter if the working model is misspecified.

6.2 Substitution

Returning to univariate estimation, we saw in Section 4 how our approach could handle parameters of the form $f(\mu_1(F_0), ..., \mu_k(F_0)) \in \mathbb{R}$, where $\mu_j(F_0)$ was a population mean. The efficient influence curve of $\mu_{n,j}$ was

$$IC_j(O|P_0) = D_j(O|G_0, Q_j(F_0)) - \mu_j(F_0).$$

We considered estimators $\mu_{n,j} = \frac{1}{n}\sum_{i=1}^n D_j(O_i|G_0, Q_j(F))$, and applied the substitution $f_n = f(\mu_{n,1}, ...\mu_{n,k})$ for the parameter of interest. The working index set is induced by full data working model $\mathcal{F}_0$ through

$$\mathcal{Q}_0 = \{(Q_1(F), ..., Q_k(F)) : F \in \mathcal{F}_0\}.$$

Using the delta method we found the influence curve

$$f_{n,(Q_1,...,Q_k)} = f(\mu_1(F_0), ..., \mu_k(F_0)) + \frac{1}{n}\sum_{i=1}^n IC(O_i|G_0, (Q_1, ..., Q_k), \eta(P_0)) + o_{P_0}(n^{-1/2}).$$

For any $(Q_1, ..., Q_k) \in \mathcal{Q}_0$, the asymptotic variance of $f_{n,(Q_1,...,Q_k)}$ was

$$\sigma^2(Q_1, ..., Q_k) = E_{P_0}[IC^2(O_i|G_0, (Q_1, ..., Q_k), \eta(P_0))].$$

Using a fit η_n for the influence curve's nuisance parameter, asymptotic variance could be approximated with the empirical mean $\frac{1}{n}\sum_{i=1}^n IC^2(O_i|G_0, (Q_1, ..., Q_k), \eta_n)$, and minimized over $\mathcal{Q}_0$. The resulting $Q_n = (Q_{n,1}, ..., Q_{n,k})$ could then be used to construct parameter estimate, and make the most efficient use of our working model $\mathcal{F}_0$ among substitution estimators $f_n = f(\mu_{n,1}, ...\mu_{n,k})$.

6.3 Estimated Coarsening Mechanism

Even when the coarsening mechanism G_0 is known, we have stressed how performance improves if we estimate it with a G_n corresponding to an efficient estimate in a correctly specified working model $\mathcal{G}_0$, and then apply $\mu_n = \frac{1}{n}\sum_{i=1}^n D(O_i|G_n, Q)$. Some may be skeptical that our technique chooses the optimal Q in working index set $\mathcal{Q}_0$ for estimating the parameter of interest. After all, we aim to minimize $\sigma^2(Q)$, or the asymptotic variance of $\frac{1}{n}\sum_{i=1}^n D(O_i|G_0, Q)$ applied with G_0 known. We could minimize a loose upper bound for asymptotic variance, and select a suboptimal $Q \in \mathcal{Q}$.

The influence curve of $\frac{1}{n}\sum_{i=1}^n D(O_i|G_n, Q)$ can be found by computing the projection in Theorem 2.3 of van der Laan and Robins (2003). It may be possible to use this influence curve to empirically approximate asymptotic variance, and jointly fit working models $\mathcal{G}_0$ and $\mathcal{Q}_0$ to produce the most efficient parameter estimate. However, our gut feeling is that this would usually be overkill.

6.4 Estimating Equations

Obviously 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. Suppose $\mu(F_0) \in \mathbb{R}$ is defined as the solution to $0 = E_{F_0}[\psi(X|\rho(F_0), \mu]$. Here $\rho(F_0)$ is a nuisance parameter. If the full data $\{X_i\}_{i=1}^n$ were available, and we could approximate the nuisance parameter at a fast enough rate with ρ_n, we could estimate $\mu(F_0)$ with the solution μ_n of $0 = \frac{1}{n}\sum_{i=1}^n \psi(X_i|\rho_n, \mu)$. Under identifiability and regularity conditions, μ_n 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 $\{O_i\}_{i=1}^n$. Applying the doubly robust mapping in their Theorem 2.1 at P_{F,G_0} (for F not necessarily equal to F_0) 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 μ_n by solving

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

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

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

As with the simpler parameters already considered, efficiently estimating the coarsening mechanism with G_n from a correctly specified submodel can only help asymptotic variance. The $Q(F)$ minimizing asymptotic variance is $Q(F_0)$, meaning it is best to apply the doubly robust mapping of full data estimating equation ψ at $P_0 = P_{F_0,G_0}$, 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 $\mathcal{Q}_0 = \{Q(F) : F \in \mathcal{F}_0\}$. We propose empirically targeting such a fit to maximize efficiency.

The influence curve's constant $c(P_0)$ is given by

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

It can be shown that this constant does not depend on $Q(F)$. We will not provide a formal proof, but the reasoning is very simple, for those acquainted with the censored data efficiency theory as presented in van der Laan and Robins (2003) or Tsiatis (2006). The expectation of $D(O|G_0, Q(F), \rho(F_0), \mu)$ can 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 $Q(F)$.

Because the constant doesn't depend on the F at which the doubly robust mapping is applied, the asymptotic variance $\sigma^2(Q)$ of the estimating equation estimator is monotone in $E_{P_0}[D^2(O_i|G_0, \rho(F_0), \mu(F_0))]$. With a coarsening mechanism estimate G_n, a preliminary parameter estimate $\hat{\mu}_n$, and a nuisance parameter estimate ρ_n, this can be approximated with $\frac{1}{n}\sum_{i=1}^{n} D^2(O_i|G_n, Q(F), \rho_n, \hat{\mu}_n)$, and minimized over a working index set $\mathcal{Q}_0$.

7 Related Methods

By no means are we the first authors to show how covariate information can guarantee asymptotic improvements over unadjusted estimators. If the unadjusted technique is an inverse probability of censoring weighted (IPCW) method such as the Horvitz-Thompson estimator, asymptotic efficiency increases when using a larger correctly specified coarsening mechanism model $\mathcal{G}_0 \subset \mathcal{G}$. Hence, while we have been discussing

the estimation problem as being defined by the choice of full data working model $\mathcal{F}_0$, another representation defines it by the size of the coarsening model $\mathcal{G}_0$ containing G_0.

Dimensionality problems can of course arise when either model becomes too large, and van der Laan and Robins (2003) recommend using relatively small working models $\mathcal{G}_0$ and $\mathcal{F}_0$. While estimation is usually straightforward in a correctly specified model $\mathcal{G}_0$, this work has focused on how to fit $F_n \in \mathcal{F}_0$. This full data distribution may sometimes be more conducive to modeling, because the $Q_n = Q(F_n)$ fit isn't used for potentially unstable inverse weighting in the resulting estimator, and consequently doesn't require artificial truncation away from zero.

Because empirical efficiency maximization differs from existing methods, several comparisons are in order.

7.1 Doubly Robust Estimation

To construct doubly robust estimates of coarsened data parameters, one fits a full data submodel with $F_n \in \mathcal{F}_0$ and a coarsening mechanism submodel with $G_n \in \mathcal{G}_0$, and uses these fits as nuisance parameter estimates in a well-chosen estimating equation for the parameter of interest. Under regularity conditions, the resulting parameter estimate will be asymptotically linear if either the $\mathcal{F}_0$ or $\mathcal{G}_0$ working model is correctly specified. This is possible because the doubly robust mapping of a full data estimating function ensures it is unbiased if either $F = F_0$ or $G = G_0$. We refer to van der Laan and Robins (2003) and the references therein for more details.

Under missingness/coarsening at random, the observed data likelihood $dP_{F,G}(O)$ factorizes into a component depending on the coarsening mechanism G, and a component depending on the full data distribution F. Maximizing likelihood in the submodel

$$\mathcal{M}_0 = \{P_{F,G} : F \in \mathcal{F}_0, G \in \mathcal{G}_0\}$$

thus splits into two separate maximizations, and a poor fit in one will not compromise the other. Hence, overviews of locally efficient estimation in van der Laan and Robins (2003) and Tsiatis (2006) have focused on fitting working models $\mathcal{G}_0$ and $\mathcal{F}_0$ with maximum likelihood, then applying these fits as nuisance parameter estimates in observed data estimating equations.

Bang and Robins (2005) make a convincing argument for doubly robust procedures in observational studies, because they give data analysts two chances for (nearly) correct model specification. Many have proposed the same likelihood based locally efficient estimators for randomized experiments, because the efficient influence curve can be identical under coarsening at random, and efficiently estimating a known G_0 can only improve asymptotic variance.

But in controlled experiments, robustness against a misspecified coarsening mechanism model $\mathcal{G}_0$ is superfluous. Right censored data has similar structure when censoring is due to the experimenter rather than the subject - essentially whenever the Kaplan-Meier estimator, or the most basic survival analysis, is justified. Empirical efficiency maximization sacrifices this unnecessary robustness for precision. We improve upon the maximum likelihood fit of the full data working model $\mathcal{F}_0$, but must generally rely on a correctly specified coarsening model to do so.

7.2 Restricted AIPWCC Estimators

Perhaps the approach bearing most resemblance to empirical efficiency maximization is the (class 1) "restricted AIPWCC" (restricted augmented inverse probability weighted complete-case) technique discussed in chapter 12 of Tsiatis (2006), and applied in survival analysis settings by Bang and Tsiatis (2000, 2002). Such estimators have beneficial properties not shared by standard locally efficient procedures, in that they can ensure efficiency gains relative to unadjusted analysis. Nevertheless, our impression is that chapter 11 of Tsiatis previews such estimators as substitutes for locally efficient estimators, to be used when more computationally feasible. This may require clarification.

Tsiatis represents the augmentation space $\{h(O) \subset L_0^2(P_0) : E[h(O)|X] = 0\}$ with a finite-dimensional basis, and elegantly derives the optimal linear combination of basis elements to augment to an inverse probability of censoring weighted estimating function, which holds for univariate or multivariate parameters. In the two-phase design, excess risk, and marginal survival curve examples considered in previous sections, this reduces to using a working index set

$$\mathcal{Q}_0 = \{\sum_{j=1}^{m} c_j Q_j : (c_1, ..., c_m) \in \mathbb{R}^m\}, \tag{15}$$

where $Q_1, ..., Q_m$ are pre-specified elements of $\mathcal{Q} = \{Q(F) : F \in \mathcal{F}\}$. Assuming a correctly specified coarsening mechanism model, Tsiatis derives the optimal linear combination $Q \in \mathcal{Q}_0$, that could be applied in a locally efficient parameter estimate.

It is clear that the representation (15) does not hold for all working index sets $\mathcal{Q}_0$ induced by a working model $\mathcal{F}_0$ in locally efficient procedures. For example, the working index set $\mathcal{Q}_0$ induced by the Cox model is not finite-dimensional. In the two-phase design problem, the logistic regression model gives rise to working index set

$$\mathcal{Q}_0 = \{Q_\beta(w) = \frac{1}{1 + \exp(-\beta^T w)} : \beta \in \mathbb{R}^p\},$$

and these functions cannot be contained in the linear span of finitely many functions. Restricted AIPWCC estimators may not be desirable in two-phase prevalence estimation, because the efficient $Q(F_0)(W) = P_{F_0}(Y = 1|W)$ has range $[0, 1]$, so wouldn't typically be represented with a basis expansion. Tsiatis et al. (2000, revised 2006) instead considered using a logistic regression working model for locally efficient estimation in developing a clinical trial strategy for the related exess risk problem, but seemingly restricted to standard maximum likelihood fits.

One can think of empirical efficiency maximization as considering a general collection of augmentation space elements, and attempting to empirically find the optimal one to augment to a univariate estimator's inverse probability weighted influence curve.

8 Discussion

While we've presented a new set of methods heuristically meliorating locally efficient estimators, much work remains. In particular, we have not formally proven claimed

asymptotic results for our estimators μ_n defined in the prevalence estimation, treatment effect, and survival analysis problems. Templates for proving asymptotic linearity are provided in the appendix, but conditions for our specific estimators have not yet been verified. We hope our covariate adjustment proposals stimulate technical empirical process analysis, in addition to computational research focusing on how to solve our new class of optimization problems.

Moreover, the i.i.d. scheme we presented for two-phase studies and randomized experiments was clearly an unrealistic approximation to how sampling would occur, and empirical efficiency maximization results must be extended beyond this setting.

In addition, we plan to supplement our Monte Carlo asymptotic efficiency calculations with simulations and data analysis. These should help gauge performance for reasonable sample sizes.

One part of our argument may appear circular or flawed. We noted efficient estimators are often eschewed because their $\sqrt{n}$-asymptotics wouldn't encapsulate performance in actual datasets, yet our subsequent analysis of locally efficient estimators was based entirely on $\sqrt{n}$-asymptotics. Our intuition was that such asymptotics could well guide performance in moderately sized samples, for a sufficiently constrained full data working model $\mathcal{F}_0$.

The method may be of interest in clinical trials, where adjustment for continuous, binary, and time-to-event outcomes is often carried out with linear, logistic, and proportional hazards models. When used in the correct intermediate step of estimator construction, and fit in a nonstandard manner, misspecified working models can guarantee increased precision from covariate information under virtually no assumptions.

We have not touched on inference. While $\frac{1}{n}\sum_{i=1}^{n} D^2(O_i|G_n, Q_n) - \mu_n^2$ may seem like a natural estimate for the asymptotic variance of our μ_n, it can be inappropriate. If the coarsening mechanism is estimated from a correct model, it will converge to asymptotic variance of the estimator applied with G_0 known, which will be an overestimate. More seriously, it could underestimate the variance of $\sqrt{n}(\mu_n - \mu(F_0))$ because of a finite sample overfitting bias. Our initial recommendation is to construct confidence intervals using the bootstrap.

In spite of the remaining work to be done, empirical efficiency maximization appears promising. In coarsened data problems where the coarsening mechanism is well understood, our method directly targets the optimal working model element for estimating a parameter of interest, enhancing locally efficient estimation, and providing a new tool for covariate adjustment in randomized experiments and survival analysis.

Appendix 1: Proofs of Theorems 1 and 2

Proof of Theorem 1. Let $Q^\star(W) = Q(F_0)(W) = E_{F_0}[Y|W]$, note that $\Delta^2 = \Delta$ as $\Delta \in \{0, 1\}$, and that $E[Y\Delta|W] = Q^\star(W)\pi_0(W)$ as Y and Δ are conditionally independent given W. We observe that,

$$\begin{aligned} \mathrm{Var}(D(O|\pi_0, Q)) &= E[D^2(O|\pi_0, Q)] - \mu^2 \\ &= E[|\frac{\Delta Y}{\pi_0(W)}|^2] + 2T_1 + T_2 - \mu^2, \end{aligned} \tag{16}$$

for

$$\begin{aligned} 2T_1 &= 2E[\frac{\Delta Y}{\pi_0(W)}(1-\frac{\Delta}{\pi_0(W)})Q(W)] \\ &= 2E[Q(W)(\frac{1}{\pi_0(W)}-\frac{1}{\pi_0^2(W)})\Delta Y] \\ &= 2E[Q(W)(\frac{1}{\pi_0(W)}-\frac{1}{\pi_0^2(W)})E[\Delta Y|W]] \\ &= 2E[Q(W)(\frac{1}{\pi_0(W)}-\frac{1}{\pi_0^2(W)})Q^\star(W)\pi_0(W)] \\ &= -2E[\frac{1-\pi_0(W)}{\pi_0(W)}Q^\star(W)Q(W)], \end{aligned} \tag{17}$$

and

$$\begin{aligned} T_2 &= E[(1-\frac{\Delta}{\pi_0(W)})^2Q^2(W)] \\ &= E[(1-2\frac{\Delta}{\pi_0^2(W)}+\frac{\Delta}{\pi_0^2(W)})Q^2(W)] \\ &= E[Q^2(W)(1-2\frac{P(\Delta=1|W)}{\pi_0(W)}+\frac{P(\Delta=1|W)}{\pi_0^2(W)})] \\ &= E[Q^2(W)(1-2+\frac{1}{\pi_0(W)}] \\ &= E[\frac{1-\pi_0(W)}{\pi_0(W)}Q^2(W)]. \end{aligned} \tag{18}$$

Combining (16), (17), and (18) we obtain,

$$\mathrm{Var}(D(O|\pi_0,Q)) = -\mu^2 + E[|\frac{\Delta Y}{\pi_0(W)}|^2] + E[\frac{1-\pi_0(W)}{\pi_0(W)}(Q^2(W)-2Q^\star(W)Q(W))].$$

The desired result follows after completing the square for $|Q^\star(W)-Q(W)|^2$, and noting that $E[|Q^\star(W)-Q(W)|^2|W] = E[|Y-Q(W)|^2|W] - E[|Y-Q^\star(W)|^2|W]$. The final step is observing $E_{P_0}[\frac{1-\pi_0(W)}{\pi_0(W)}|Y-Q(W)|^2] = E_{P_0}[\Delta\frac{1-\pi_0(W)}{\pi_0^2(W)}|Y-Q(W)|^2]$, as it is easy to check that $E_{P_0}[\frac{\Delta}{\pi_0(W)}\psi(W,Y)] = E_{P_0}[\psi(W,Y)]$ for any integrable $\psi(W,Y)$. □

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,

$$
\begin{aligned}
& E_{P_0}\left\{\frac{\Delta I(\tilde{T}>t)}{\bar{G}_0(\tilde{T}_-)}\right\}\left\{\int \frac{Q(c,W)}{\bar{G}_0(c)}dM(c)\right\} \\
&= E_{P_0}\left\{\frac{\Delta I(\tilde{T}>t)}{\bar{G}_0(\tilde{T}_-)}\right\}\left\{\int \frac{Q(c,W)}{\bar{G}_0(c)}(dN(c)-dA(c))\right\} \\
&= -E_{P_0}\left\{\frac{\Delta I(\tilde{T}>t)}{\bar{G}_0(\tilde{T}_-)}\right\}\left\{\int \frac{Q(c,W)}{\bar{G}_0(c)}dA(c)\right\} \\
&= -E_{P_0}\left\{\frac{\Delta I(\tilde{T}>t)}{\bar{G}_0(\tilde{T}_-)}\right\}\left\{\int \frac{Q(c,W)I(\tilde{T}\geq c)}{\bar{G}_0(c)\bar{G}_0(c_-)}dG_0(c)\right\} \\
&= -E_{P_0}\int \frac{\Delta I(\tilde{T}>t)}{\bar{G}_0(\tilde{T}_-)}\frac{Q(c,W)I(\tilde{T}\geq c)}{\bar{G}_0(c)\bar{G}_0(c_-)}dG_0(c) \\
&= -E_{P_0}\int \beta(\tilde{T},\Delta,c)Q(c,W)dG(c) = -\int\{\beta(\tilde{t},\delta,c)Q(c,w)\}dG_0(c)dP_0(w,\delta,\tilde{t}) \\
&= -E_{(O,C)\sim P_0\times G_0}[\beta(\Delta,\tilde{T},C)Q(C,W)] \qquad (19)
\end{aligned}
$$

Further, standard martingale results (i.e. Theorem 2.6.1 in Fleming and Harrington (1991)), imply

$$
\begin{aligned}
E_{P_0}|\int \frac{Q(c,W)}{\bar{G}_0(c)}dM(c)|^2 &= E_{P_0}\int |\frac{Q(c,W)}{\bar{G}_0(c)}|^2\phi(c,\tilde{T})dA(c) \\
&= E_{P_0}\int |\frac{Q(c,W)}{\bar{G}_0(c)}|^2\phi(c,\tilde{T})I(\tilde{T}\geq c)\frac{dG_0(c)}{\bar{G}_0(c_-)} \\
&= E_{P_0}\int \alpha(\tilde{T},c)Q^2(c,W)dG_0(c) \\
&= \int \alpha(\tilde{t},c)Q^2(c,w)dG_0(c)dP_0(w,\tilde{t}) \\
&= E_{(O,C)\sim P_0\times G_0}[\alpha(\tilde{T},C)Q^2(C,W)] \qquad (20)
\end{aligned}
$$

Squaring the sum $D(O|G_0,Q)$ and finding the expectation of the three terms with (19) and (20) yields the desired result. □

Appendix 2: Proving Asymptotic Linearity

In this appendix, we'll provide guidelines for proving empirical efficiency maximization leads to an asymptotically linear estimator. Proofs might appear very familiar. They should have the same structure as proofs for standard locally efficient estimators, where the working index set $\mathcal{Q}_0$ corresponds to a working model $\mathcal{F}_0$ for the data generating distribution, and is fit by maximum likelihood. The only difference is our new objective function, as we try to choose the $F_n \in \mathcal{F}_0$ minimizing the empirical mean of $D^2(O|G_0,Q(F))$ rather $-\log dP_{F,G_0}(O)$.

Let M_n denote the stochastic process indexed by working index set $\mathcal{D}_0 = \{D(\cdot|G_0, Q) : Q \in \mathcal{Q}_0\}$, given by $M_n(D) = \mathbb{P}_n D^2(O)$, and let $M : \mathcal{D}_0 \to \mathbb{R}$ denote the deterministic function $M(D) = P_0 D^2(O)$. Let the nuisance parameter estimate $Q_n \in \mathcal{Q}_0$ be an empirical efficiency maximizer satisfying,

$$\frac{1}{n}\sum_{i=1}^{n} D^2(O_i|G_n, Q_n) \leq \inf_{D\in\mathcal{D}_0} M_n(D) + o_{P_0}(1). \tag{21}$$

Note that this should require the $G_n \to G_0$ convergence. We consider applying the estimator $\mathbb{P}_n\{D(\cdot|G_n, Q_n)\}$, and would like to prove its asymptotic linearity, meaning

$$\mathbb{P}_n\{D(\cdot|G_n, Q_n)\} = \frac{1}{n}\sum_{i=1}^{n} IC(O_i|P_0) + o_{P_0}(n^{-1/2}) \text{ for } IC(\cdot|P_0) \in L_0^2(P_0).$$

We will define $D(\cdot|G_0, Q_0) = \text{argmin}_{D\in\mathcal{D}_0} M(D)$ as the oracle's element of $\mathcal{D}_0$ for estimation of $\mu = \mu(F_0)$, and

(A) assume the minimizer $D(\cdot|G_0, Q_0)$ of $E_{P_0}[D^2(O|G_0, Q)]$ exists and is P_0-unique.

Note the uniqueness refers to the function $D(\cdot|G_0, Q_0)$, and not to elements of working index set $\mathcal{Q}_0$.

Our first step is to show our estimator is asymptotically equivalent to the estimator we could construct if we knew the oracle's nuisance parameter Q_0, but still estimated the coarsening mechanism G_0 with G_n. That is,

$$\begin{aligned}\mathbb{P}_n\{D(\cdot|G_n, Q_n) - D(\cdot|G_n, Q_0)\} &= (\mathbb{P}_n - P_0)\{D(\cdot|G_n, Q_n) - D(\cdot|G_n, Q_0)\} \\ &+ P_0\{D(O|G_n, Q_n) - D(O|G_n, Q_0)\} \\ &= o_{P_0}(n^{-1/2}). \end{aligned} \tag{22}$$

To prove this asymptotic equivalence (22), we will likely need tools from empirical process theory. We refer to van der Vaart and Wellner (1996) as a reference, particularly for the formal framework in which the forthcoming statements should be understood. A useful well-known result from empirical process theory is that if f_n is a sequence of functions, possibly randomly determined by the data $\{O_i\}_{i=1}^n$, that $(\mathbb{P}_n - P_0)\{f_n(\cdot)\} = o_{P_0}(n^{-1/2})$ if $P_0 f_n^2(O) \to 0$ in probability and there is a P_0-Donsker class containing f_n with probability tending to one. In light of this fact, it is immediate that (22) is satisfied under the following assumptions:

(B1) $P_0\{D(O|G_n, Q_n) - D(O|G_n, Q_0)\} = o_{P_0}(n^{-1/2})$.

(B2) There exists a subset $\mathcal{G}_0 \subset \mathcal{G}$ containing G_0, and also containing G_n with probability tending to one, such that,

$$\{D(\cdot|G, Q) : G \in \mathcal{G}_0, Q \in \mathcal{Q}_0\} \text{ is a } P_0\text{-Donsker class of functions.}$$

(B3) $P_0\{D(O|G_n, Q_n) - D(O|G_n, Q_0)\}^2 \to 0$ in probability.

When will these assumptions be satisfied? (B1) will depend more on the parametric $G_n \to G_0$ convergence rather than the $Q_n \to Q_0$ convergence. Note that if G_0 is known and we take $G_n = G_0$, then the condition will be satisfied with the right side equal to zero, as $E_{P_0}[D(O|G_0, Q)] = \mu$ for all $Q \in \mathcal{Q}_0$.

Condition (B2) restricts the amount of nuisance parameters G and Q we're allowed to consider when constructing our estimator. If the coarsening mechanism G_0 is known and will be used in the estimator $\mathbb{P}_n\{D(\cdot|G_0, Q_n)\}$, then we can take $\mathcal{G}_0 = \{G_0\}$. Verification should be straightforward when working index set $\mathcal{Q}_0$ corresponds to a familiar parametric or semiparametric working model such as a logistic regression or Cox model.

To show (B3), first consider the case of G_0 known, and examine the $D(\cdot|G_0, Q_n) \to D(\cdot|G_0, Q_0)$ convergence. (When G_0 is unknown but $\mathcal{Q}_0$ is a sufficiently restricted working index set, we will typically only additionally need the consistency of G_n for G_0 in some sense.) Endow $\mathcal{D}_0 = \{D(\cdot|G_0, Q) : Q \in \mathcal{Q}_0\}$ with the distance metric $d(D_1, D_2) = \sqrt{P_0|D_1 - D_2|^2}$. Corollary 3.2.3(ii) of van der Vaart and Wellner (1996) suffices under condition (A) if:

(C1) $\sup_{D \in K} |M_n(D) - M(D)| \to 0$ in probability for every compact $K \subset \mathcal{D}_0$,

(C2) The map $D \to M(D)$ is lower semicontinuous.

(C3) The sequence $D_n = D(\cdot|G_0, Q_n)$ is uniformly tight.

(C1) is a restriction on the size of our working index set $\mathcal{Q}_0$, as we will just need for $\{D^2(\cdot|G_0, Q) : Q \in \mathcal{Q}_0\}$ to be a Glivenko-Cantelli class of functions. Like (B2), this is usually no trouble to prove when estimating nuisance parameters using well-known parametric or semiparametric working models. (C2) is an analytic rather than probabilistic condition that we must prove by examining $M(D) = P_0 D^2(O)$.

After we've used (B1)-(B3) and (C1)-(C3) to establish asymptotic equivalence of $\mathbb{P}_n\{D(\cdot|G_n, Q_n)\}$ and $\mathbb{P}_n\{D(\cdot|G_n, Q_0)\}$ as in (22), there are two cases to consider. The first is when the coarsening mechanism G_0 is exactly known, and we take $G_n = G_0$. In such a setting, (22) tells us our estimator $\mathbb{P}_n\{D(\cdot|G_n, Q_n)\}$ is asymptotically linear with influence curve,

$$IC_0(O|P_0) = D(O|G_0, Q_0) - \mu(F_0).$$

That is, our estimator and the oracle's estimator are asymptotically equivalent. The second case is when G_0 isn't known, but G_n is an asymptotically efficient estimate in our model $\mathcal{M}$, as in marginal survival curve estimation with independent censoring. In this situation, we can write,

$$\begin{aligned} \mathbb{P}_n\{D(\cdot|G_n, Q_0)\} - \mu(F_0) &= (\mathbb{P}_n - P_0)\{D(\cdot|G_n, Q_0) - D(\cdot|G_0, Q_0)\} \\ &+ \frac{1}{n}\sum_{i=1}^{n} IC_0(O_i|P_0) \\ &+ R_n(G_n) \end{aligned}$$

for

$$R_n(G_n) = P_0\{D(O|G_n, Q_0) - D(O|G_0, Q_0)\}.$$

Our previously mentioned empirical process result for showing $(\mathbb{P}_n - P_0)\{f_n(\cdot)\} = o_{P_0}(n^{-1/2})$ implies this can be reduced to

$$\mathbb{P}_n\{D(\cdot|G_n, Q_0)\} - \mu = \frac{1}{n}\sum_{i=1}^{n} IC_0(O_i|P_0) + R_n(G_n) + o_{P_0}(n^{-1/2}) \tag{23}$$

if the Donsker condition (B2) holds and we have the convergence,

(D) $P_0\left\{D(O|G_n, Q_0) - D(O|G_0, Q_0)\right\}^2 \to 0$ in probability.

Verifying (D) should only require showing $G_n \to G_0$ in some sense, and not a rate or efficiency result for this convergence. The remaining task is analyzing $R_n(G_n)$. When G_n is efficient for G in the model $\mathcal{M}$, it should follow that:

(E1) There is an $IC_{\text{nuis}}(\cdot|P_0)$ with mean zero and finite variance under P_0 such that

$$R_n(G_n) = \frac{1}{n}\sum_{i=1}^{n} IC_{\text{nuis}}(O_i|P_0) + o_{P_0}(n^{-1/2}).$$

(E2) The function $IC_{\text{nuis}}(\cdot|P_0)$ belongs to the tangent space of the model $\mathcal{M}$ at P_0, and is orthogonal in $L_0^2(P_0)$ to $IC_{\text{efficient}}(\cdot|P_0)$, the efficient influence curve (canonical gradient) for estimation of $\mu(F_0)$. That is,

$$< IC_{\text{nuis}}(\cdot|P_0), IC_{\text{efficient}}(\cdot|P_0) >_{L_0^2(P_0)} = E_{P_0}[IC_{\text{nuis}}(O|P_0)IC_{\text{efficient}}(O|P_0)] = 0.$$

We refer to Bickel, Klassen, Ritov, and Wellner (1998) for an overview of relevant semiparametric theory, and formal definitions for the tangent space and efficient influence curve. In words, the tangent space is the linear closure in the Hilbert space $L_0^2(P_0)$ of the span of scores of regular parametric submodels of $\mathcal{M}$ passing through P_0. The efficient influence curve is a scaled version of the efficient score, or score of the regular parametric submodel of $\mathcal{M}$ through P_0 in which estimation of parameter $\mu = \mu(F_0)$ is most difficult in terms of an information bound.

Condition (E1) implies that our estimator $\mathbb{P}_n\{D(\cdot|G_n, Q_n)\}$ is an asymptotically linear estimator of parameter $\mu = \mu(F_0)$ at P_0, with influence curve,

$$IC(O|P_0) = IC_0(O|P_0) + IC_{\text{nuis}}(O|P_0).$$

To demonstrate (E1), we will once again usually need that $G_n \to G_0$ in some sense at the parametric $n^{-1/2}$ rate. One way to find $IC_{\text{nuis}}(\cdot|P_0)$ might be to first find the influence curve of G_n as an estimator of G_0, which could be infinite dimensional, and then apply the functional delta method to the function $G \to P_0\{D(\cdot|G, Q_0)\}$.

After finding $IC_{\text{nuis}}(\cdot|P_0)$, we can check the stronger condition (E2). This will typically require G_n to not only approach G_0 at the $n^{-1/2}$ rate, but to be efficient for G_0 in the coarsening mechanism model $\mathcal{G}_0$. The benefit of (E2) is that it implies the influence curve $IC(O|P_0)$ has variance no larger than that of $IC_0(O|P_0)$. This means our estimator built from nuisance parameters G_n and Q_n is as asymptotically efficient as the estimator $\mathbb{P}_n\{D(\cdot|G_0, Q_0)\}$ we could use if knowing the coarsening mechanism

G_0 and the oracle's nuisance parameter $Q_0 \in \mathcal{Q}_0$. For a formal justification of this well-known but somewhat paradoxical result that efficiently estimating a known coarsening mechanism can only improve an estimator's asymptotics, we refer to Theorem 2.3 of van der Laan and Robins (2003).

To summarize, we propose using the estimator $\mathbb{P}_n\{D(\cdot|G_n, Q_n)\}$, where Q_n is the empirical efficiency maximizer satisfying (21). Under checkable conditions, this estimator is asymptotically linear and at least as asymptotically efficient as the oracle estimator $\mathbb{P}_n\{D(\cdot|G_0, Q_0)\}$. Existence and uniqueness of this efficiency maximizing $Q_0 \in \mathcal{Q}_0$ are assumed in (A). (C1)-(C3) are useful for demonstrating a $Q_n \to Q_0$ convergence to prove (B1)-(B3), and (B1)-(B3) imply $\mathbb{P}_n\{D(\cdot|G_n, Q_n)\}$ and $\mathbb{P}_n\{D(\cdot|G_n, Q_0)\}$ are asymptotically equivalent as in (22). If using $G_n = G_0$, this result (22) reveals asymptotic equivalence with $\mathbb{P}_n\{D(\cdot|G_0, Q_0)\}$. If G_n is instead an efficient estimate of the coarsening mechanism G_0, our estimator is still at least this efficient under (E1), (E2), and (23), where (23) follows from (B2) and (D).

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] Bang, H. and Robins, J.M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61, 962-973.

[3] Bang, H. and Tsiatis, A.A. (2000). Estimating medical costs with censored data. *Biometrika*, 87, 329-343.

[4] Bang, H. and Tsiatis, A.A. (2002). Median regression with censored cost data. *Biometrics*, 58, 643-649.

[5] 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.

[6] Breslow, N.E. (1974). Covariance analysis of censored survival data. *Biometrics.* 30, 89-99.

[7] 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.

[8] Cox, D.R. (1972). Regression models and life tables (with discussion). *Journal of the Royal Statistical Society B,* 34, 187-220.

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

[10] Fleming, T.R. and Harrington, D.P. (1991). *Counting Processes and Survival Analysis*. John Wiley & Sons, Inc.

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

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

[13] Friedman, L.M., Furberg, C.D., and DeMets, D.L. (1998). *Fundamentals of Clinical Trials* (third edition). Springer-Verglag, New York.

[14] 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.

[15] Heitjan, D.F. and Rubin, D.B. (1991). Ignorability and coarse data. *Annals of Statistics*, 19, 2244-2253.

[16] 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.

[17] Kaplan, E.L., and P. Meier. (1958). Nonparametric estimator from incomplete observations. *Journal of the American Statistical Association.* 53, 457-481.

[18] 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.

[19] Lu, X. and Tsiatis, A.A. (2007). Improving the efficiency of the logrank test using auxiliary covariates. *Biometrika*, in press.

[20] Mantel, N. (1984). Pre-stratification or post-stratification. *Biometrics*, 40, 256-258 (letter).

[21] 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.

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

[23] Neyman, J. (1938). Contribution to the theory of sampling human populations. *Journal of the American Statistical Association.* 33, 101-116.

[24] Peto, R. (1978). Clinical trail methodology. *Biomedicine*, 28, 24-36 (special issue).

[25] 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.

[26] Robins, J.M. and Ritov, Y. (1997). Toward a curse of dimensionality appropriate (CODA) asymptotic theory for semi-parametric models. *Statistics in Medicine,* 16, 285-319.

[27] 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.

[28] 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.

[29] 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.

[30] Robinson, L.D. and Jewell, N.P. (1991). Some surprising results about covariate adjustment in logistic regression. *International Statistical Review*, 59, 227-240.

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

[32] 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.

[33] 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.

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

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

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

[37] van der Laan, M.J. and Rubin, D.B. (2006). Targeted maximum likelihood learning. *The International Journal of Biostatistics*, Volume 2, Issue 1, Article 11.

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

[39] Zeng, D. (2004). Estimating marginal survival function by adjusting for dependent censoring using many covariates. *Annals of Statistics.* Vol. 32, No. 4, 1533-1555.