Abstract

LUNCEFORD, JARED KENNETH. Estimating Causal Treatment Effects Via the Propensity Score and Estimating Survival Distributions in Clinical Trials That Follow Two-stage Randomization Designs. (Under the direction of Professor Marie Davidian)

Estimation of treatment effects with causal interpretation from obervational data is complicated by the fact that exposure to treatment is confounded with subject characteristics. The propensity score, the probability of exposure to treatment conditional on covariates, is the basis for two competing classes of approaches for adjusting for confounding: methods based on stratification of observations by quantiles of estimated propensity scores, and methods based on weighting individual observations by weights depending on estimated propensity scores. We review these approaches and investigate their relative performance.

Some clinical trials follow a design in which patients are randomized to a primary therapy upon entry followed by another randomization to maintenance therapy contingent upon disease remission. Ideally, analysis would allow different treatment policies, i.e. combinations of primary and maintenance therapy if specified up-front, to be compared. Standard practice is to conduct separate analyses for the primary and follow-up treatments, which does not address this issue directly. We propose consistent estimators of the survival distribution and mean survival time for each treatment policy in such two-stage studies and derive large sample properties. The methods are demonstrated on a leukemia clinical trial data set and through simulation.

ESTIMATING CAUSAL TREATMENT EFFECTS VIA THE PROPENSITY SCORE AND ESTIMATING SURVIVAL DISTRIBUTIONS IN CLINICAL TRIALS THAT FOLLOW TWO-STAGE RANDOMIZATION DESIGNS

by

JARED KENNETH LUNCEFORD

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

STATISTICS

Raleigh

2001

APPROVED BY:

Marie Davidian Chair of Advisory Committee	Anastasios Tsiatis
Dennis Boos	Daowen Zhang

For my family

Personal Biography

Jared Kenneth Lunceford was born in Provo, Utah on July 20, 1972. He was brought up under the care of outstanding parents—Kenneth Charles Lunceford and Annette Hancey Lunceford. While in Utah he met a lovely girl, Heidi Jasperson. They were married on December 9, 1989 and have shared a wonderful life together since.

Jared obtained his undergraduate education at Brigham Young University, completing his B.S. in Molecular Biology in August 1994. He decided to pursue a doctoral degree in the same; however, during his graduate studies he became interested in the field of statistics. He finished an M.S. in Molecular Biology at Brigham Young University and, with his family, left Utah to enter the statistics graduate program at North Carolina State University in August of 1996. After spending five years in beautiful North Carolina completing Jared's doctoral degree, Jared, Heidi, and their three sons (Janceton, Ashton, and Mason) will move to New Jersey; there Jared will join the Clinical Biostatistics department of Merck & Company.

Acknowledgements

I am very glad to have occasion to write this acknowledgments section. It affords me the opportunity to voice my gratitude to those persons who have helped me achieve this milestone in my education. I feel their influence is longlasting and that I and those dear to me will continue to enjoy the fruits of that influence.

Direct thanks for the help and guidance that produced this dissertation and the training involved goes to my advisor, Dr. Marie Davidian. I feel most fortunate to have worked with so talented a person these past years. She always goes the extra mile in her work and has been a great example to me professionally. Much of the knowledge and skills I gained at this university have come through my association with her during research and class work. I have great admiration for her strong work ethic, organization, and intellect and have benefitted from such. I would also like to give her my thanks on behalf of my family.

I would like to thank Dr. Anastasios Tsiatis. He was also very involved in my research, and I appreciate the time he spent working with me in that regard. I would like him to know how much I enjoyed attending his lectures; that I always looked forward to taking his courses. Some of the most advanced material I encountered during my time here was also some of the most understandable material, largely due to his skills and enthusiasm.

At the risk of sounding a little immodest, I will say that my decision to train in this department was a very good one—for I have had a uniformly positive experience.

Overall I think this is due to a fabulous group of faculty and staff. I would like to convey my appreciation to Dr. Bibhuti Bhattacharyya and Dr. Dennis Boos for their work as instructors during my Measure Theory and Advanced Inference sequences. Their doors were always open for a visit, and they always patiently entertained my questions during class. I especially appreciate Dr. Bhattacharyya's willingness to hold extra study sessions and to continue with lectures on advanced probability topics when I and a few others requested such. I would like to thank Dr. Sastry Pantula who always accommodated me at those junctures when I wanted to change course. I would like to thank him for the level of financial support I received. My thanks also go to Dr. Bruce Weir in that regard for his support during my first years here. Thanks to Dr. Boos and Dr. Daowen Zhang for their service on my graduate committee. Thanks also to Janice Gaddy and Sharon Patton for their sweet kindness and willingness to assist in any way.

I would like to thank a very important friend, Jimmy Doi. He has been a true friend to me, Heidi, and our sons. It has been such a blessing to pass through the rigors of this program in his ever pleasant and considerate company. Not only that, he never seems to mind buying lunch three times for every one lunch that I buy.

On behalf of me and my wife, I would like to thank our wonderful parents for their support. Special thanks go to my father and mother for their financial assistance during our stay in North Carolina. I know there are many thanks they deserve that I have yet to recognize, but with regard to completing this doctoral degree, I want to convey my gratitude for the priority they placed on my performance in the public

schools and their unconditional support of my graduate efforts.

I can not properly express the affection and gratitude I feel towards my wife, Heidi. There is no official diploma awaiting her for over a decade of selfless service to our family while I have been in school, but I hope it is sufficient to say that I dearly love her, that I thank God for her in my prayers, and that my greatest joy at this time is the peace and happiness we share at home.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Stratification and Weighting Via the Propensity Score in Estimation of Causal Treatment Effects

Observational data are often the basis for epidemiological and other investigations where the objective is to make inference on the effect of treatment exposure on a response. Randomized studies serve to balance distributions of subject characteristics across groups, so that groups are similar except for the treatments. However, with observational data, treatment exposure may be associated with covariates that are also associated with potential response, and groups may be seriously imbalanced in these factors. Consequently, unbiased comparisons of treatments from observational data require methods that adjust for likely confounding of exposure to treatment

with subject characteristics, and inferences on effects of treatments with a causal interpretation cannot be made without appropriate adjustment.

For a study comparing two treatments, "treated" and "control," say, the *propensity score* is the probability of exposure to treatment conditional on a set of observed covariates (Rosenbaum and Rubin 1983). Properties of the propensity score that facilitate such causal inferences are elucidated by Rosenbaum and Rubin (1983) (for an introduction, see D'Agostino 1998 and Rosenbaum 1998), and applications of methods that use adjustments based on propensity scores are increasingly widespread (e.g. Connors et al. 1996; Shepardson et al. 1999; Perkins et al. 2000). The most popular method for estimating the (causal) difference of two treatment means is that of Rosenbaum and Rubin (1984), where individuals are stratified based on estimated propensity scores and the difference is estimated as the average of within-stratum effects. An alternative approach is to adjust for confounding by using estimated propensity scores to construct weights for individual observations (e.g. Rosenbaum 1987; Robins, Hernán, and Brumback 2000).

Stratification and weighting based on propensity scores are two general approaches to making causal statements from observational data. The first half of this dissertation investigates their relative performance. In Chapter 2, we review the framework of counterfactuals or potential outcomes (e.g. Rubin 1974), which formalizes the notion of a "causal effect" and assumptions required to justify adjustments for confounding, and propensity-score based estimators. We review several estimators of causal treatment effects that employ the propensity score. Chapter 3 outlines large-sample

properties, and Chapter 4 reports results from the theory and simulations that provide insight into how the estimators compare under different conditions.

1.2 Estimation of Survival Distributions of Treatment Policies in Two-Stage Randomization Designs in Clinical Trials

Cancer therapy frequently is implemented via a two-stage approach, where an initial treatment is given with the intent of inducing disease remission, and a follow-up treatment is given to prolong the period before relapse and disease progression. In many cases, this second-stage maintenance therapy is given only to those patients who show a complete or partial remission in response to the induction therapy. Similarly, HIV patients may receive an initial therapy, which may be modified contingent on response to the initial treatment.

Cancer clinical trials studying combinations of induction and maintenance therapies are common (e.g., Stein et al., 1990; Joss et al., 1994; Tummarello et al., 1997). After enrollment, patients are randomized to an induction therapy, followed by a subsequent randomization to maintenance therapy contingent on their remission status and consent. Protocol 8923, conducted by the Cancer and Leukemia Group B (CALGB) and reported by Stone et al. (1995), examined the effects of infusions of granulocyte-macrophage colony-stimulating factor (GM-CSF) after ini-

tial chemotherapy in 388 elderly patients with acute myelogenous leukemia (AML). Standard chemotherapy for AML has a myelosuppressive effect, placing patients at increased risk of death due to infection or bleeding-related complications. As a hematopoietic cytokine, GM-CSF administered after chemotherapy might assist patient recovery by allowing a more rapid reconstitution of bone-marrow-derived lineages, thus reducing the number of deaths due to chemotherapy-related complications. CALGB 8923 was a double-blind, placebo-controlled, two-stage trial in which patients were initially randomized to GM-CSF or placebo following standard chemotherapy. Later, patients meeting the complete remission criteria and consenting to further participation were then randomized to one of two intensification treatments.

A main interest is to compare treatment policies, i.e. induction/maintenance therapy combinations if offered to patients up-front, and to determine the combination leading to the greatest survival benefit. As is customary, the analysis would focus on comparing treatment policies under the "intent-to-treat" principle. However, like the randomization scheme used in these trials, data analysis typically is separated into two parts, neither of which addresses this issue directly: (i) estimating survival distribution under different induction therapies using all data while ignoring maintenance therapy, and (ii) estimating post-remission survival distribution using only data for individuals receiving maintenance therapy. We believe that this is a consequence of lack of available methodology. In the second half of this dissertation, we propose a framework in which estimation of survival distributions and other quantities relevant to comparing treatment policies with an "intent-to-treat" interpretation may be car-

ried out. The methods are simple to implement and offer an unambiguous approach to estimating treatment policy effects and possible interactions between induction and maintenance therapies. Chapter 5 introduces the counterfactual framework and the proposed estimators of survival. The empirical properties are investigated in Chapter 6, along with a demonstration of the methods using a leukemia clinical trial data set.

Chapter 2

Estimators of Causal Effects Based on the Propensity Score

2.1 Counterfactual Framework

Consider two treatments, and let Y be the response, Z be an indicator of treatment exposure ($Z = 1$ if treated, $Z = 0$ if control), and $\boldsymbol{X}$ be a vector of covariates, so that $(Y, Z, \boldsymbol{X})$ are observed on each individual. The elements of $\boldsymbol{X}$ are assumed measured prior to receipt of treatment or, if measured post-treatment, are not affected by either treatment, e.g. $\boldsymbol{X}$ may include baseline covariates or demographic characteristics. Each individual is assumed to have an associated random vector (Y_0, Y_1), where Y_0 and Y_1 are the responses that would be observed if, possibly contrary to fact, s/he received control or treatment, respectively. Consequently, Y_0 and Y_1 are referred to as *counterfactuals* (or *potential outcomes*) and may be viewed as inherent characteristics

of the individual, the values of response that would be seen if an individual were treated or given control, usually formalized by assuming that Y satisfies

$$Y = Y_1 Z + (1 - Z)Y_0. \tag{2.1}$$

It is important to distinguish between the observed response Y and the counterfactuals: Y_0 and Y_1 are hypothetical and may never be observed simultaneously; however, they are a convenient construct that allows precise statement of questions of interest.

The distributions of Y_0 and Y_1 may be thought of as representing the hypothetical distributions of response for the population of individuals were all individuals to receive control or be treated, respectively, so the means of these distributions correspond to the mean response if all individuals were to receive each treatment. Hence, a difference in these means would be attributable to, or caused by, the treatments; formally, then,

$$\Delta = \mu_1 - \mu_0 = E(Y_1) - E(Y_0)$$

is referred to as the average causal effect (of the treated state relative to control). Estimation of Δ is thus of central interest in comparing the treatments.

Within this framework, it is possible to state formally the difficulty in estimating Δ, and thus making causal statements, from observational data. The counterfactuals are never both observed for any subject; thus, whether estimation of Δ is possible relies on whether $E(Y_0)$ and $E(Y_1)$ may be identified from the observed data $(Y, Z, \boldsymbol{X})$. In an observational study, because treatment exposure is not controlled, treatment status Z may not be independent of the counterfactuals (Y_0, Y_1); indeed, the same

characteristics that lead an individual to be exposed to a treatment may also be associated with his/her potential response. Under these conditions and (2.1), the usual sample average in the treated group estimates

$$E(Y|Z=1) = E(Y_1|Z=1) \neq E(Y_1) \tag{2.2}$$

and similarly for the control group, so that the difference of observed sample averages is not an unbiased estimator for Δ. In contrast, in a randomized trial, as Z is determined for each participant at random, it is unrelated to how s/he might respond, and thus $(Y_0, Y_1) \perp\!\!\!\perp Z$, where $\perp\!\!\!\perp$ denotes statistical independence. Under this condition, the inequality in (2.2) becomes an equality, and the difference in observed sample averages is indeed an unbiased estimator of Δ with a causal interpretation, as is widely accepted.

In an observational study, although $(Y_0, Y_1) \perp\!\!\!\perp Z$ is unlikely to hold, it may be possible to identify subject characteristics $\boldsymbol{X}$, sometimes called "confounders", that are related to potential response and treatment exposure. If we further believe that $\boldsymbol{X}$ contains all such characteristics that determine treatment exposure, then, for all individuals with a particular value of $\boldsymbol{X}$, there would be no association between the exposure states and the values of the counterfactuals; i.e. treatment exposure among individuals with a particular $\boldsymbol{X}$ is essentially at random. Formally, the counterfactuals are independent of treatment exposure conditional on $\boldsymbol{X}$, written

$$(Y_0, Y_1) \perp\!\!\!\perp Z \,|\, \boldsymbol{X}. \tag{2.3}$$

Rosenbaum and Rubin (1983) refer to (2.3) as the assumption of strongly ignorable

treatment assignment; (2.3) has also been called the assumption of no unmeasured confounders (see Robins, Hernán, and Brumback 2000). One must appreciate that (2.3) is an *assumption*; willingness to assume (2.3) requires the analyst to have confidence that $\boldsymbol{X}$ contains all characteristics related to determination of treatment and that there are no additional, unmeasured such factors.

The benefit of (2.3) is that $E(Y_0)$ and $E(Y_1)$ may be identified from $(Y, Z, \boldsymbol{X})$. For given $\boldsymbol{X}$, $E(Y|Z = z, \boldsymbol{X})$, $z = 0, 1$, depends only on the observed data, so is identifiable. But, taking $z = 1$ as an example, its average over all $\boldsymbol{X}$ satisfies $E\{\, E(Y|Z = 1, \boldsymbol{X})\,\} = E\{\, E(Y_1|Z = 1, \boldsymbol{X})\,\} = E\{\, E(Y_1|\boldsymbol{X})\,\} = E(Y_1)$, where the first equality is from (2.1) and the second follows from strong ignorability (2.3); similarly, $E\{\, E(Y|Z = 0, \boldsymbol{X})\,\} = E(Y_0)$. This demonstrates that it is possible to make inferences on Δ from observational data when (2.3) may be assumed to hold. Methods using the propensity score are one way to achieve this.

The propensity score $e(\boldsymbol{X}) = P(Z = 1|\boldsymbol{X})$, $0 < e(\boldsymbol{X}) < 1$, is the probability of treatment given the observed covariates. Rosenbaum and Rubin (1983) showed that $\boldsymbol{X} \perp\!\!\!\perp Z \,|\, e(\boldsymbol{X})$, so individuals from either treatment group with the same propensity score are balanced in the sense that the distribution of $\boldsymbol{X}$ is the same regardless of treatment status. Rosenbaum and Rubin in fact show that if (2.3) holds, in addition $(Y_0, Y_1) \perp\!\!\!\perp Z \,|\, e(\boldsymbol{X})$, so that treatment exposure is unrelated to the counterfactuals for individuals sharing the same propensity score. We now review ways these properties may be exploited to derive estimators for Δ from $(Y_i, Z_i, \boldsymbol{X}_i)$, $i = 1, \ldots, n$, an i.i.d. sample containing both treated and control subjects.

2.2 Estimation of the Propensity Score

In practice, the propensity score is unlikely to be known, so it is routine to estimate it from the observed data $(Z_i, \boldsymbol{X}_i)$, $i = 1, \ldots, n$, by assuming that $e(\boldsymbol{X})$ follows a parametric model, e.g. a logistic regression model $e(\boldsymbol{X}, \boldsymbol{\beta}) = \{1 + \exp(-\boldsymbol{X}^T\boldsymbol{\beta})\}^{-1}$ where $\boldsymbol{\beta}$ is $(p \times 1)$; interaction and higher-order terms may also be included. From the $(Z_i, \boldsymbol{X}_i)$, $\boldsymbol{\beta}$ may be estimated by standard techniques, e.g. by the maximum likelihood estimator $\widehat{\boldsymbol{\beta}}$, which we assume in the sequel, solving in $\boldsymbol{\beta}$

$$\sum_{i=1}^{n} \psi_\beta(Z_i, \boldsymbol{X}_i, \boldsymbol{\beta}) = \sum_{i=1}^{n} \frac{Z_i - e(\boldsymbol{X}_i, \boldsymbol{\beta})}{e(\boldsymbol{X}_i, \boldsymbol{\beta})\{1 - e(\boldsymbol{X}_i, \boldsymbol{\beta})\}} \partial/\partial\boldsymbol{\beta}\{e(\boldsymbol{X}_i, \boldsymbol{\beta})\} = 0. \tag{2.4}$$

We henceforth assume that the analyst is proficient at modeling the propensity score, so that $e(\boldsymbol{X}, \boldsymbol{\beta})$ is correctly specified, and for brevity we write $e = e(\boldsymbol{X}, \boldsymbol{\beta})$ and $e_\beta = \partial/\partial\boldsymbol{\beta}\{e(\boldsymbol{X}, \boldsymbol{\beta})\}$, with subscript i when evaluated at $\boldsymbol{X}_i$.

2.3 Estimation of Δ Based on Stratification

The popular approach using stratification on estimated propensity scores to estimate Δ involves the following general steps: (i) Estimate $\boldsymbol{\beta}$ as in (2.4) and calculate estimated propensity scores $\widehat{e}_i = e(\boldsymbol{X}_i, \widehat{\boldsymbol{\beta}})$ for all i; (ii) form K strata according to the sample quantiles of the $\widehat{e}_i$, where the jth sample quantile $\widehat{q}_j$, $j = 1, \ldots, K$, is such that the proportion of $\widehat{e}_i \leq \widehat{q}_j$ is roughly j/K, $\widehat{q}_0 = 0$, and $\widehat{q}_K = 1$; (iii) within each stratum, calculate the sample mean of the Y_i for each treatment; and (iv) estimate $E(Y_1)$ and $E(Y_0)$ by a weighted sum of the sample means in (iii) across strata, where

weighting is by the proportion of observations falling in its stratum, and estimate Δ by difference. Formally, the resulting estimator is

$$\begin{aligned}\widehat{\Delta}_S &= \sum_{j=1}^{K}\left(\frac{n_j}{n}\right) n_{1j}^{-1}\sum_{i=1}^{n} Z_i Y_i I(\widehat{e}_i \in \widehat{Q}_j) \\ &\quad -\sum_{j=1}^{K}\left(\frac{n_j}{n}\right)(n_j - n_{1j})^{-1}\sum_{i=1}^{n}(1-Z_i)Y_i I(\widehat{e}_i \in \widehat{Q}_j), \qquad (2.5)\end{aligned}$$

where $\widehat{Q}_j = (\widehat{q}_{j-1}, \widehat{q}_j]$, $n_j = \sum_{i=1}^{n} I(\widehat{e}_i \in \widehat{Q}_j)$ is the number of individuals in stratum j, and $n_{1j} = \sum_{i=1}^{n} Z_i I(\widehat{e}_i \in \widehat{Q}_j)$ is the number of these who are treated.

The rationale follows from the property $(Y_0, Y_1) \perp\!\!\!\perp Z \,|\, e(\boldsymbol{X})$ when (2.3) holds; because treatment exposure is essentially at random for individuals with the same propensity value, we expect mean comparisons within this group to be unbiased. Identifying individuals sharing exactly the same propensity value may be infeasible in practice, so stratification attempts to achieve groups where this at least holds approximately. Consequently, $\widehat{\Delta}_S$ may be a biased estimator of Δ, as some residual confounding within strata may remain. Rosenbaum and Rubin (1983, 1984) advocate the use of quintiles ($K = 5$) following results of Cochran (1968) that suggest this remaining bias may be small, to remove up to 90% made, a choice made in most published applications. Intuitively, these results require that the propensity model be correctly specified. Thus, it is often recommended (e.g. Perkins et al. 2000, Rosenbaum and Rubin 1984) that, following (ii), the analyst examine the degree of balance for each element of $\boldsymbol{X}$ within each stratum using standard statistical tests. Evidence that balance has not been achieved may reflect an incorrect model and the need for refinement, following by a return to (i).

For theoretical arguments (Chapter 3), it proves convenient to rewrite (2.5). Replacing n_j/n with its limit $1/K$ and writing $\widehat{p}_j = n_{1j}/n$, an asymptotically equivalent version is

$$\widehat{\Delta}_S = n^{-1}\sum_{i=1}^{n}\frac{Z_iY_i}{K}\left\{\sum_{j=1}^{K}\frac{I(\widehat{e}_i \in \widehat{Q}_j)}{\widehat{p}_j}\right\} - n^{-1}\sum_{i=1}^{n}\frac{(1-Z_i)Y_i}{K}\left\{\sum_{j=1}^{K}\frac{I(\widehat{e}_i \in \widehat{Q}_j)}{1/K-\widehat{p}_j}\right\}. \quad (2.6)$$

2.4 Estimation of Δ Based on Weighting

Rather than seek to achieve unbiased estimation within strata, weighting methods attempt to obtain an unbiased estimator for Δ directly in a way akin to that proposed by Horvitz and Thompson (1952) in the context of sampling with unequal selection probabilities. Under (2.1), $E\{ZY/e(\boldsymbol{X})\} = E\{ZY_1/e(\boldsymbol{X})\}$, so that, assuming (2.3) and noting $Z = I(Z = 1)$,

$$\begin{aligned} E\left\{\frac{ZY}{e(\boldsymbol{X})}\right\} &= E\left[E\left\{\frac{I(Z=1)Y_1}{e(\boldsymbol{X})}\middle| Y_1, \boldsymbol{X}\right\}\right] \\ &= E\left[\frac{Y_1}{e(\boldsymbol{X})}E\{I(Z=1)\,|\,Y_1,\boldsymbol{X}\}\right] \\ &= E(Y_1), \end{aligned}$$

where (2.3) allows the last equality; similarly, $E[(1-Z)Y/\{1-e(\boldsymbol{X})\}] = E(Y_0)$. These results suggest immediately the estimator for Δ proposed by Rosenbaum (see Rosenbaum 1998) and others

$$\widehat{\Delta}_{IPW1} = n^{-1}\sum_{i=1}^{n}\frac{Z_iY_i}{\widehat{e}_i} - n^{-1}\sum_{i=1}^{n}\frac{(1-Z_i)Y_i}{1-\widehat{e}_i}. \quad (2.7)$$

An alternative is suggested by $E\{Z/e(\boldsymbol{X})\} = E\{\, E(Z \mid \boldsymbol{X})/e(\boldsymbol{X})\} = 1$ and similarly $E[(1-Z)/\{1-e(\boldsymbol{X})\}] = 1$:

$$\widehat{\Delta}_{IPW2} = \left(\sum_{i=1}^{n} \frac{Z_i}{\widehat{e}_i}\right)^{-1} \sum_{i=1}^{n} \frac{Z_i Y_i}{\widehat{e}_i} - \left(\sum_{i=1}^{n} \frac{1-Z_i}{1-\widehat{e}_i}\right)^{-1} \sum_{i=1}^{n} \frac{(1-Z_i)Y_i}{1-\widehat{e}_i}. \tag{2.8}$$

The estimator for a single mean in (2.8), e.g. μ_1, also referred to as a ratio estimator in the sampling literature, may be written as the solution to $\sum_{i=1}^{n} Z_i(Y_i - \mu_1)/\widehat{e}_i = 0$, while the usual sample mean solves $\sum_{i=1}^{n}(Y_i - \mu_1) = 0$; the single-mean estimators in (2.7) may be written similarly. Thus, "IPW" stands for "inverse-probability-of-complete-case-weighting" of estimating equations for a mean as we now describe.

In particular, $\widehat{\Delta}_{IPW1}$ and $\widehat{\Delta}_{IPW2}$ may be deduced within the semiparametric missing data framework of Robins, Rotnitzky, and Zhao (1994). To appreciate the connection with missing data problems, consider μ_1; identifying $(Y_1, Z, \boldsymbol{X})$ to be the "full data," Y_1 is only observed for individuals with $Z = 1$, so that the probability of a "complete case" is $P(Z = 1|\boldsymbol{X})$ if treatment is related to $\boldsymbol{X}$. Under these conditions, the work of Robins et al. (1994) characterizes the class of all semiparametric estimators for μ_1 and μ_0; i.e. estimators under the condition that the distributions of Y_1, Y_0, and $\boldsymbol{X}$ are unspecified. Estimators in this class are consistent if the complete-case probability is correctly modeled and hence should be approximately unbiased in finite samples. The class includes simple estimators using inverse-probability-of-complete-case weighting, as in $\widehat{\Delta}_{IPW1}$ and $\widehat{\Delta}_{IPW2}$ [for μ_0, the complete-case probability is $P(Z = 0|\boldsymbol{X}) = 1 - P(Z = 1|\boldsymbol{X})$], but others are more complex. The work of Robins et al. (1994) identifies the estimator within the class having the smallest

(large-sample) variance, the semiparametric efficient estimator

$$\begin{aligned}\widehat{\Delta}_{DR} &= n^{-1}\sum_{i=1}^{n}\frac{Z_iY_i-(Z_i-\widehat{e}_i)m(1,\boldsymbol{X}_i,\widehat{\boldsymbol{\alpha}})}{\widehat{e}_i} \\ &\quad -n^{-1}\sum_{i=1}^{n}\frac{(1-Z_i)Y_i+(Z_i-\widehat{e}_i)m(0,\boldsymbol{X}_i,\widehat{\boldsymbol{\alpha}})}{1-\widehat{e}_i}. \qquad (2.9)\end{aligned}$$

Here $m(z,\boldsymbol{X},\boldsymbol{\alpha}) = E(Y|Z=z,\boldsymbol{X})$ is the regression of the response on Z and $\boldsymbol{X}$, depending on parameters $\boldsymbol{\alpha}$, and $\widehat{\boldsymbol{\alpha}}$ is an estimator for $\boldsymbol{\alpha}$. Unlike $\widehat{\Delta}_S$, $\widehat{\Delta}_{IPW1}$, and $\widehat{\Delta}_{IPW2}$, $\widehat{\Delta}_{DR}$ requires specification of this regression model. However, as pointed out by Scharfstein, Rotnitzky, and Robins (1999, sec. 3.2.3), $\widehat{\Delta}_{DR}$ has a so-called "doubly-robust" property that the estimator remains consistent even if either the propensity score model e or the regression model m (but not both) is incorrectly specified. Neither $\widehat{\Delta}_{IPW1}$ nor $\widehat{\Delta}_{IPW2}$ would be expected to be consistent if e were incorrectly specified, as the motivating arguments at the beginning of this section would no longer be valid. As we focus on estimation based on the propensity score only, we view $\widehat{\Delta}_{DR}$ as a benchmark for comparison when both $e(\boldsymbol{X},\boldsymbol{\beta})$ and $m(z,\boldsymbol{X},\boldsymbol{\alpha})$ are correctly chosen.

Because it attempts to use information on the distribution of $e(\boldsymbol{X})$ (in the form of classification based on sample quantiles), $\widehat{\Delta}_S$ is not a member of this class of semiparametric estimators and hence cannot be represented in the form of such estimators. Thus, insights into its properties relative to those of $\widehat{\Delta}_{IPW1}$, $\widehat{\Delta}_{IPW2}$, and $\widehat{\Delta}_{DR}$ are not immediate from the Robins et al. (1994) theory.

An alternative to all estimators previously discussed is estimation of Δ from a regression model. This is possible directly for the linear model $E(Y|Z=z,\boldsymbol{X}) =$

$m(z, \boldsymbol{X}, \boldsymbol{\alpha}) = \alpha_0 + \alpha_1 z + \alpha_2^T \boldsymbol{X}$; under (2.3), it is straightforward to verify that $E\{E(Y|Z = 1, \boldsymbol{X})\} - E\{E(Y|Z = 0, \boldsymbol{X})\} = \alpha_1 = \Delta$. For nonlinear models, this difference does not have a closed form, and integration of $m(z, \boldsymbol{X}, \boldsymbol{\alpha})$ over an estimate of the distribution of $\boldsymbol{X}$ is required. For either linear or nonlinear models, a potential drawback is that, when the dimension of $\boldsymbol{X}$ is large, ensuring that the model is correct is difficult. Although a similar issue arises for modeling propensity score, because this approach does not directly model covariate effects on the counterfactual means, it has been argued that the approach may be less susceptible to erroneous inferences on Δ.

Chapter 3

Large Sample Properties of Estimators of Average Causal Effect

3.1 Weighted Estimators

The properties of $\widehat{\Delta}_{IPW1}$, $\widehat{\Delta}_{IPW2}$, and $\widehat{\Delta}_{DR}$ when e is correctly specified and $\boldsymbol{\beta}$ is estimated may be deduced by writing them as the solutions to estimating equations of the form $\sum_{i=1}^{n} \psi_{\Delta}(Y_i, Z_i, \boldsymbol{X}_i, \Delta, \boldsymbol{\beta}) = 0$ solved jointly with (2.4). For $\widehat{\Delta}_{DR}$, we also assume that the regression function is correctly specified; here, ψ_{Δ} also depends on $\boldsymbol{\alpha}$, which is estimated by solving an additional equation of form $\sum_{i=1}^{n} \psi_{\alpha}(Y_i, Z_i, \boldsymbol{X}_i, \boldsymbol{\alpha}) = \mathbf{0}$, e.g., as for least squares or logistic regression. Letting Δ_0 denote the true value of Δ, the theory of Robins et al. (1994) shows that each estimator is such that

$n^{1/2}(\widehat{\Delta} - \Delta_0)$ converges in distribution to a mean-zero normal random variable with variance Σ. Because the results follow straightforwardly from Robins al. (1994), we forego explicit derivation of these variances.

For $\widehat{\Delta}_{IPW1}$ and $\widehat{\Delta}_{IPW2}$, if $\boldsymbol{\beta}$ were known instead of estimated, so that $e(\boldsymbol{X}, \boldsymbol{\beta})$ is a known function of $\boldsymbol{X}$, the large-sample variances are

$$\Sigma^*_{IPW1} = E\left(\frac{Y_1^2}{e} + \frac{Y_0^2}{1-e}\right) - \Delta_0^2 \quad \text{and} \quad \Sigma^*_{IPW2} = E\left\{\frac{(Y_1-\mu_1)^2}{e} + \frac{(Y_0-\mu_0)^2}{1-e}\right\},$$

where the expectations are with respect to the joint distribution of $(Y_0, Y_1, \boldsymbol{X})$ and all parameters are equal to their true values. If, as in practice, $\boldsymbol{\beta}$ is estimated, then the variances become, with $\boldsymbol{E}_{\beta\beta} = E[e_\beta e_\beta^T/\{e(1-e)\}]$,

$$\Sigma_{IPW1} = \Sigma^*_{IPW1} - \boldsymbol{H}^T_{\beta,1}\boldsymbol{E}^{-1}_{\beta\beta}\boldsymbol{H}_{\beta,1}, \quad \boldsymbol{H}_{\beta,1} = E\left\{\left(\frac{Y_1}{e} + \frac{Y_0}{1-e}\right)e_\beta\right\}, \tag{3.1}$$

$$\Sigma_{IPW2} = \Sigma^*_{IPW2} - \boldsymbol{H}^T_{\beta,2}\boldsymbol{E}^{-1}_{\beta\beta}\boldsymbol{H}_{\beta,2}, \quad \boldsymbol{H}_{\beta,2} = E\left\{\left(\frac{Y_1-\mu_1}{e} + \frac{Y_0-\mu_0}{1-e}\right)e_\beta\right\}, \tag{3.2}$$

thus exhibiting the interesting property that estimating $\boldsymbol{\beta}$ rather than knowing its value leads to a reduction in variance in both estimators. However, there is no general ordering of Σ_{IPW1} and Σ_{IPW2}, so that neither estimator is uniformly more precise (asymptotically). For $\widehat{\Delta}_{DR}$, it may be shown from Robins et al. (1994) that the variance is

$$\Sigma_{DR} = \Sigma^*_{IPW2} - E\left[\sqrt{\frac{1-e}{e}}\{E(Y_1|\boldsymbol{X}) - \mu_1\} + \sqrt{\frac{e}{1-e}}\{E(Y_0|\boldsymbol{X}) - \mu_0\}\right]^2. \tag{3.3}$$

The Robins et al. theory guarantees that $\Sigma_{DR} \leq \Sigma_{IPW1}$ and Σ_{IPW2}. As long as the propensity and regression models do not share parameters, Σ_{DR} is the same whether

$\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ are known or estimated. The forms of (3.1)–(3.3) depend on quantities that may be estimated from the data, so large sample approximations to the sampling variances of the estimators may be derived readily.

3.2 Stratification Estimator

The properties of $\widehat{\Delta}_S$ are more difficult and to our knowledge have not been elucidated; in practice, an ad hoc estimator of sampling variance is used (see Section 4.1). We now present a heuristic sketch of the argument based on representing the stratification scheme as the solution to a set of estimating equations. For technical reasons, we take $\boldsymbol{X}$ to be continuous, although this could be relaxed with more rigorous arguments. As in practice, we take the number of strata K to be predetermined and fixed.

Unlike $\widehat{\Delta}_{IPW1}$ and $\widehat{\Delta}_{IPW2}$, $\widehat{\Delta}_S$ involves estimation not only of $\boldsymbol{\beta}$ but also of the true quantiles $\boldsymbol{q} = (q_1, \ldots, q_{K-1})^T$ and the probabilities $\boldsymbol{p} = (p_1, \ldots, p_K)^T$ that an individual is treated and has propensity score in $Q_j = (q_{j-1}, q_j]$, where $q_0 = 0$, $q_K = 1$. Estimation of $\boldsymbol{\beta}$ is by solving (2.4), as before. Equations corresponding to estimation of p_j and q_j have forms

$$\sum_{i=1}^n \psi^S_{p_j}(Z_i, \boldsymbol{X}_i, q_{j-1}, q_j, p_j, \boldsymbol{\beta}) = \sum_{i=1}^n Z_i I(e_i \in Q_j) - p_j = 0, \quad j = 1, \ldots, K,$$
$$\sum_{i=1}^n \psi^S_{q_j}(\boldsymbol{X}_i, q_j, \boldsymbol{\beta}) = \sum_{i=1}^n I(e_i \leq q_j) - j/K = 0, \quad j = 1, 2, ...K-1,$$

respectively. The latter equations do not have zero solutions for some n, but by taking $\widehat{q}_j$ to be the inverse of the empirical CDF for the $\widehat{e}_i$ are $o(n^{1/2})$ at the solutions, so we downplay this technicality in the following. Even with $e(\boldsymbol{X}, \boldsymbol{\beta})$ correctly specified, as

noted in Section 2.3, we expect $\widehat{\Delta}_S$ to be inconsistent due to failure of stratification to eliminate all confounding. In particular, from (2.6), given that all elements of $\boldsymbol{\theta} = (\boldsymbol{q}^T, \boldsymbol{p}^T, \boldsymbol{\beta}^T)^T$ are estimated consistently, by the law of large numbers $\widehat{\Delta}_S$ converges in probability to $\Delta^* = \mu_1^* - \mu_0^*$, where $\mu_1^* = K^{-1}\sum_{j=1}^K E\{Y_1 eI(e \in Q_j)\}/E\{eI(e \in Q_j)\}$, and $\mu_0^* = K^{-1}\sum_{j=1}^K E\{Y_0(1-e)I(e \in Q_j)\}/[K^{-1} - E\{eI(e \in Q_j)\}]$. Note that, by an argument similar to that leading to (2.8), a sufficient condition for $\Delta^* = \Delta_0$ is $(Y_0, Y_1) \perp\!\!\!\perp \boldsymbol{X}$, in which case confounding is not an issue, as would be expected, but, in general, $\Delta^* \neq \Delta_0$. Thus, $\widehat{\Delta}_S$ estimates Δ^*, so from (2.6)

$$\begin{aligned}\sum_{i=1}^n \psi^S_{\Delta^*}(Y_i, Z_i, \boldsymbol{X}, \Delta^*, \boldsymbol{\theta}) &= \sum_{i=1}^n \left\{ Z_i Y_i K^{-1} \sum_{j=1}^K I(e_i \in Q_j)/p_j \right. \\ &\qquad \left. -(1-Z_i) Y_i K^{-1} \sum_{j=1}^K I(e_i \in Q_j)/(k^{-1} - p_j) - \Delta^* \right\} \\ &= 0.\end{aligned}$$

is an equation corresponding to estimation of Δ based on stratification. Writing $\Psi_\theta = (\psi^S_{q_1}, \ldots, \psi^S_{q_{K-1}}, \psi^S_{p_1}, \ldots, \psi^S_{p_K}, \psi_\beta)^T$, then, $\widehat{\Delta}_S$ and $\widehat{\boldsymbol{\theta}}$ solve

$$\sum_{i=1}^n \{\Psi_\theta^T(Z_i, \boldsymbol{X}_i, \boldsymbol{\theta}), \psi^S_{\Delta^*}(Y_i, Z_i, \boldsymbol{X}, \Delta^*, \boldsymbol{\theta})\}^T = \mathbf{0}. \tag{3.4}$$

The properties of $\widehat{\Delta}_S$ may be derived from (3.4) by appealing to standard arguments (e.g. Carroll, Ruppert, and Stefanski, 1995, sec. A.3.6), under assumptions that allow these to be augmented to account for the nondifferentiability of some elements of (3.4) in q_j and $\boldsymbol{\beta}$ (e.g. Huber, 1981, Chap. 3). These arguments show that $n^{1/2}(\widehat{\Delta}_S - \Delta^*)$ converges in distribution to a mean-zero normal random variable with

variance

$$\Sigma_S = \Sigma_S^* + \Gamma_p + \Gamma_{qp} + \Gamma_{\beta qp} \tag{3.5}$$

for Γ_p, Γ_{qp}, and $\Gamma_{\beta qp}$ given in Appendix A. In (3.5), letting $f_e(\cdot)$ be the density of the propensity scores and $E(\cdot|e)$ be conditional expectation given the propensity score,

$$\begin{aligned}\Sigma_S^* &= K^{-2}\sum_{j=1}^{K} p_j^{-2}\int_{q_{j-1}}^{q_j} E(Y_1^2|t)tf_e(t)dt \\ &\quad + K^{-2}\sum_{j=1}^{K}(1/K - p_j)^{-2}\int_{q_{j-1}}^{q_j} E(Y_0^2|t)(1-t)f_e(t)dt - (\Delta^*)^2\end{aligned}$$

and equals the variance of $\widehat{\Delta}_S$ we would obtain if the q_j, p_j, and $\boldsymbol{\beta}$ were all known. The remaining terms represent additional effects of estimating $\boldsymbol{q}$, $\boldsymbol{p}$, and $\boldsymbol{\beta}$ given that the remaining terms in the subscript have also been estimated; e.g. $\Gamma_{\beta qp}$ is the effect of estimating $\boldsymbol{\beta}$ rather than knowing it if $\boldsymbol{q}$ and $\boldsymbol{p}$ are estimated. Inspection of Σ_S^* and the expressions in Appendix A shows that, not unexpectedly and unlike $\widehat{\Delta}_{IPW1}$ and $\widehat{\Delta}_{IPW2}$, the large-sample variance of $\widehat{\Delta}_S$ depends critically on the densities of the propensity score and $\boldsymbol{X}$.

The form of (3.5) is obviously not attractive for derivation of an approximate sampling variance, as it is both complex and would require estimation of the densities involved. However, comparing Σ_S to Σ_{IPW1}, Σ_{IPW2}, and Σ_{DR} indicates clearly that the properties of the two types of estimators for Δ may be expected to differ, and whether one type outperforms the other in general is not immediately apparent. In contrast to the situation in (3.1) and (3.2), it is not possible to deduce that any of the terms Γ_p, Γ_{qp}, or $\Gamma_{\beta qp}$ are negative, which would imply that estimation of these parameters reduces variance relative to the (unlikely) situation where they would be

known. Interestingly, it may be shown that $\Gamma_{\beta qp} = 0$ in the event $\boldsymbol{X}$ is a scalar and $e(\boldsymbol{X}, \boldsymbol{\beta})$ is the simple logistic regression model.

Availability of the expression Σ_S does allow comparison of large-sample properties under specific scenarios, as we demonstrate in Section 4.1. The nature of $\Gamma_{\beta qp}$ for the case of two, bivariate-normal confounders is investigated empirically in Section 4.2.

3.3 Effect of Additional Covariates

In the previous development, it has been assumed that $\boldsymbol{X}$ is associated with both treatment exposure and potential response and that (2.3) holds. For $\widehat{\Delta}_S$, a common guideline is that it is preferable to "over-model" the propensity score by including additional covariates that do not affect treatment exposure rather than run the risk of excluding relevant ones (e.g. McIntosh and Rubin 1999; Perkins et al. 2000). In fact, intuition would suggest that including such covariates when they are correlated with potential response could provide additional information on Δ. It is possible to formalize this situation as follows.

Suppose $\boldsymbol{V}$ is an additional set of covariates, exclusive of $\boldsymbol{X}$, that (i) does not affect treatment exposure but (ii) is associated with potential response. More precisely, (i) may be written as $P(Z = 1|\boldsymbol{X}, \boldsymbol{V}) = P(Z = 1|\boldsymbol{X})$, and (ii) implies that the conditional distributions of Y_0 and Y_1 given $(\boldsymbol{X}, Z, \boldsymbol{V})$ depend on $\boldsymbol{V}$. Suppose that the analyst is willing to assume strong ignorability given both $\boldsymbol{X}$ and $\boldsymbol{V}$, i.e.

$$(Y_0, Y_1) \perp\!\!\!\perp Z \,|\, (\boldsymbol{X}, \boldsymbol{V}). \tag{3.6}$$

It is straightforward to show using manipulations similar to those in Dawid (1979) that (3.6) implies that (2.3) also holds. Thus, it is possible to specify a model $P(Z = 1|\boldsymbol{X}, \boldsymbol{V}) = e(\boldsymbol{X}, \boldsymbol{V}, \boldsymbol{\beta}, \boldsymbol{\gamma})$, where $\boldsymbol{\gamma}$ is an additional $(q \times 1)$ parameter corresponding to terms in the model involving $\boldsymbol{V}$, such that this model reduces to the true propensity score $e(\boldsymbol{X}, \boldsymbol{\beta})$ (depending on $\boldsymbol{X}$ and $\boldsymbol{\beta}$ only) when $\boldsymbol{\gamma} = \mathbf{0}$, its "true" value, and the assumptions underlying the derivations of (3.1)– (3.3) and (3.5) hold. Suppose, then, that the chosen propensity score model satisfies $e(\boldsymbol{X}, \boldsymbol{V}, \boldsymbol{\beta}, \mathbf{0}) = e(\boldsymbol{X}, \boldsymbol{\beta}) = e$ and is such that $\partial/\partial\boldsymbol{\beta}\{e(\boldsymbol{X}, \boldsymbol{V}, \boldsymbol{\beta}, \boldsymbol{\gamma})\}|_{\gamma=0} = e_\beta$ depending on $\boldsymbol{X}$ and $\boldsymbol{\beta}$ only; e.g. $e(\boldsymbol{X}, \boldsymbol{V}, \boldsymbol{\beta}, \boldsymbol{\gamma}) = [1 + \exp\{-(\boldsymbol{X}^T\boldsymbol{\beta} + \boldsymbol{V}^T\boldsymbol{\gamma})\}]^{-1}$.

Under these circumstances, for any of the methods, Δ will be estimated jointly with both the previous additional parameters and $\boldsymbol{\gamma}$. The effect of including $\boldsymbol{V}$ in the propensity score model may thus be deduced by considering the previous estimating equations for each estimator, replacing $e(\boldsymbol{X}, \boldsymbol{\beta})$ by $e(\boldsymbol{X}, \boldsymbol{V}, \boldsymbol{\beta}, \boldsymbol{\gamma})$, and adding the additional equation

$$\begin{aligned}
&\sum_{i=1}^{n} \psi_\gamma(Z_i, \boldsymbol{X}_i, \boldsymbol{V}_i\boldsymbol{\beta}, \boldsymbol{\gamma}) \\
&= \sum_{i=1}^{n} \frac{Z_i - e(\boldsymbol{X}_i, \boldsymbol{V}_i, \boldsymbol{\beta}, \boldsymbol{\gamma})}{e(\boldsymbol{X}_i, \boldsymbol{V}_i, \boldsymbol{\beta}, \boldsymbol{\gamma})\{1 - e(\boldsymbol{X}_i, \boldsymbol{V}_i, \boldsymbol{\beta}, \boldsymbol{\gamma})\}} \partial/\partial\boldsymbol{\gamma}\{e(\boldsymbol{X}_i, \boldsymbol{V}_i, \boldsymbol{\beta}, \boldsymbol{\gamma})\} \\
&= \mathbf{0}. \qquad (3.7)
\end{aligned}$$

Note that $\partial/\partial\boldsymbol{\gamma}\{e(\boldsymbol{X}, \boldsymbol{V}, \boldsymbol{\beta}, \boldsymbol{\gamma})\}$ evaluated at the "truth" $\boldsymbol{\gamma} = \mathbf{0}$ may depend on both $\boldsymbol{X}$ and $\boldsymbol{V}$; in the above example, this partial derivative equals $\boldsymbol{V}/[e(\boldsymbol{X}, \boldsymbol{V}, \boldsymbol{\beta}, \boldsymbol{\gamma})\{1 - e(\boldsymbol{X}, \boldsymbol{V}, \boldsymbol{\beta}, \boldsymbol{\gamma})\}]$. In general, write $e_\gamma = \partial/\partial\boldsymbol{\gamma}\{e(\boldsymbol{X}, \boldsymbol{V}, \boldsymbol{\beta}, \boldsymbol{\gamma})\}|_{\gamma=\mathbf{0}}$, with subscript i when evaluated at $(\boldsymbol{X}_i, \boldsymbol{V}_i)$.

From these augmented estimating equations for each estimator, it is possible to derive properties under (3.6), e.g. by exploiting Section A.3.6 of Carroll et al. (1995). The details are tedious but straightforward; a sketch for $\widehat{\Delta}_{IPW1}$ is given in Appendix A. Defining $\boldsymbol{E}_{\gamma\gamma} = E[e_\gamma e_\gamma^T/\{e(1-e)\}]$ and $\boldsymbol{E}_{\gamma\beta} = E[e_\gamma e_\beta^T/\{e(1-e)\}]$, and letting $\boldsymbol{H}_{\gamma\beta} = \boldsymbol{E}_{\gamma\gamma} - \boldsymbol{E}_{\gamma\beta}\boldsymbol{E}_{\beta\beta}^{-1}\boldsymbol{E}_{\gamma\beta}^T$, all weighted estimators still are such that $n^{1/2}(\widehat{\Delta} - \Delta_0)$ converges in distribution to a mean-zero normal random variable, now with variance Σ^V. For $\widehat{\Delta}_{IPW1}$ and $\widehat{\Delta}_{IPW2}$, we have

$$\Sigma_{IPW1}^V = \Sigma_{IPW1} - (\boldsymbol{H}_{\gamma,1} - \boldsymbol{E}_{\gamma\beta}\boldsymbol{E}_{\beta\beta}^{-1}\boldsymbol{H}_{\beta,1})^T\boldsymbol{H}_{\gamma\beta}^{-1}(\boldsymbol{H}_{\gamma,1} - \boldsymbol{E}_{\gamma\beta}\boldsymbol{E}_{\beta\beta}^{-1}\boldsymbol{H}_{\beta,1}), \qquad (3.8)$$

$$\Sigma_{IPW2}^V = \Sigma_{IPW2} - (\boldsymbol{H}_{\gamma,2} - \boldsymbol{E}_{\gamma\beta}\boldsymbol{E}_{\beta\beta}^{-1}\boldsymbol{H}_{\beta,2})^T\boldsymbol{H}_{\gamma\beta}^{-1}(\boldsymbol{H}_{\gamma,2} - \boldsymbol{E}_{\gamma\beta}\boldsymbol{E}_{\beta\beta}^{-1}\boldsymbol{H}_{\beta,2}), \qquad (3.9)$$

$$\boldsymbol{H}_{\gamma,1} = E\left\{\left(\frac{Y_1}{e} + \frac{Y_0}{1-e}\right)e_\gamma\right\}, \quad \boldsymbol{H}_{\gamma,2} = E\left\{\left(\frac{Y_1-\mu_1}{e} + \frac{Y_0-\mu_0}{1-e}\right)e_\gamma\right\}.$$

From (3.8) and (3.9), the effect of including $\boldsymbol{V}$ in the propensity score model is to reduce the variance relative to that in the case where $\boldsymbol{V}$ is excluded. The practical implication is that, at least in large samples, incorporating a covariate that does not affect treatment exposure into the propensity model will always lead to greater precision for estimating Δ using these weighted estimators in the event the covariate is related to response.

With the addition of $\boldsymbol{V}$, the form of the semiparametric efficient estimator is modified from (2.9). The semiparametric efficient estimator is now that with the smallest large-sample variance among all semiparametric estimators for which the

distributions of Y_0, Y_1, $\boldsymbol{X}$, and $\boldsymbol{V}$ are unspecified. In particular,

$$\begin{aligned}\widehat{\Delta}^V_{DR} &= n^{-1}\sum_{i=1}^{n}\frac{Z_iY_i-(Z_i-\widehat{\widehat{e}}_i)m(1,\boldsymbol{X}_i,\boldsymbol{V}_i,\widehat{\boldsymbol{\delta}})}{\widehat{\widehat{e}}_i}\\ &\quad -n^{-1}\sum_{i=1}^{n}\frac{(1-Z_i)Y_i+(Z_i-\widehat{\widehat{e}}_i)m(0,\boldsymbol{X}_i,\boldsymbol{V}_i,\widehat{\boldsymbol{\delta}})}{1-\widehat{\widehat{e}}_i},\end{aligned} \tag{3.10}$$

where now $\widehat{\widehat{e}}_i = e(\boldsymbol{X}_i,\boldsymbol{V}_i,\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{\gamma}})$, and $m(z,\boldsymbol{X},\boldsymbol{V},\delta) = E(Y|Z=z,\boldsymbol{X},\boldsymbol{V})$ is the regression of Y on Z, $\boldsymbol{X}$, and $\boldsymbol{V}$ depending on parameters $\boldsymbol{\delta}$ estimated by $\widehat{\boldsymbol{\delta}}$. As before, this estimator requires modeling of the regression and maintains the "doubly-robust" property. It may be shown from the Robins et al. (1994) theory that the variance of $\widehat{\Delta}^V_{DR}$ is given by

$$\Sigma^V_{DR} = \Sigma^*_{IPW2} - E\left[\sqrt{\frac{1-e}{e}}\{E(Y_1|\boldsymbol{X},\boldsymbol{V})-\mu_1\}+\sqrt{\frac{e}{1-e}}\{E(Y_0|\boldsymbol{X},\boldsymbol{V})-\mu_0\}\right]^2 \tag{3.11}$$

and satisfies $\Sigma^V_{DR} \leq \Sigma_{DR}$; of course $\Sigma^V_{DR} \leq \Sigma^V_{IPW1}$ and Σ^V_{IPW2}. Comparing (3.3) and (3.11), it is clear that a gain in efficiency is possible by incorporating the information on potential response in $\boldsymbol{V}$ for this estimator as well.

As $e(\boldsymbol{X},\boldsymbol{V},\boldsymbol{\beta},\boldsymbol{0}) = e$, the stratification estimator will still converge in probability to Δ^*; however, the variance of $n^{1/2}(\widehat{\Delta}_S - \Delta^*)$ will change. Applying the same arguments to the stratification scheme, where now (3.7) is solved jointly with the previous equations, yields variance

$$\Sigma^V_S = \Sigma_S + \Gamma_{\gamma\beta qp}, \tag{3.12}$$

where $\Gamma_{\gamma\beta qp}$ is given in Appendix A. Unfortunately, in contrast to the results for the weighted estimators above, the sign of $\Gamma_{\gamma\beta qp}$ is not immediately evident, so it is not

clear whether incorporation of $\boldsymbol{V}$ need lead to a reduction in variance; if $\Gamma_{\gamma\beta qp} > 0$, the situation could worsen. Obviously, the properties of the stratification estimator are considerably more complex. In Section 4.3, we investigate this issue empirically.

Chapter 4

Empirical Results for Estimators of Average Causal Effect

4.1 Univariate Confounder

As a simple scenario under which the asymptotic performance of these estimators can be investigated, consider the case where treatment is conditionally independent of the counterfactuals given the value of a single normally distributed scalar X with propensity model $e(X) = \{1 + \exp(-\beta_0 - \beta_1 X)\}^{-1}$. Results are presented for both binary and continuous responses having regression models $E(Y|X, Z) = \{1 + \exp(-\alpha_0 - \alpha_1 X - \alpha_2 Z)\}^{-1}$ and $E(Y|X, Z) = \alpha_0 + \alpha_1 X + \alpha_2 Z$, respectively. In this and the following sections we compare asymptotic relative mean squared error,

denoted ARE, between $\widehat{\Delta}_S$ and the various weighted estimators using

$$\text{ARE}_m = \frac{(\Delta^* - \Delta)^2 + n^{-1}\Sigma_S}{n^{-1}\Sigma_m},$$

where $m = IPW1, IPW2$, or DR.

Exact and Monte Carlo ARE results are shown for a 2×2 combination of α_1 and β_1 values. As is common in practice, five strata were used in the computation of $\widehat{\Delta}_S$. The values of β_0, α_0, and α_1 were fixed at 0, 0, and 1, respectively. In all cases the confounder X is distributed $N(2, 1)$. For generating the response in the continuous case, the conditional distribution of Y given X and Z was $N(\alpha_0 + \alpha_1 X + \alpha_2 Z, 1)$.

Table 4.1 shows exact ARE results in the case where β_1 is estimated. Figures 4.1 and 4.2 show the ARE profiles generated from the numbers given in Table 4.1 as n increases. A prominent feature of these figures is the important role of β_1 in the asymptotic relative efficiency of weighting to stratification—the larger the magnitude of β_1, the more the predicted decrease in efficiency of the weighted estimators; an effect that is magnified to some extent by the strength of the association between X and prognosis through the value of α_1. Not surprisingly, the large sample bias of $\widehat{\Delta}_S$ dominates as sample size increases.

Before discussing Monte Carlo results, we note that using the findings of Section 3.2 to obtain standard errors for $\widehat{\Delta}_S$ is not practical. In practice, it is routine to approximate the sampling variance of $\widehat{\Delta}_S$ by treating $\widehat{\Delta}_S$ as the average of K, independent, within stratum, treatment effect estimates. Assuming an equal number

of individuals per strata, standard errors for $\widehat{\Delta}_S$ are computed using

$$K^{-2}\sum_{j=1}^{K}\hat{\sigma}_j^2 \tag{4.1}$$

where $\hat{\sigma}_j^2$ is an estimate of the variance of the difference between the treatment means in stratum j. Using the notation of Section 2.3, $\hat{\sigma}_j^2 = n_{1j}^{-1}s_{1j}^2 + (n_j - n_{1j})^{-1}s_{0j}^2$, where $s_{1j}^2 = n_{1j}^{-1}\sum_{i=1}^{n} I(\widehat{e}_i \in \widehat{Q}_j)(Z_iY_i - \overline{y}_{1j})^2$, $s_{0j}^2 = (n_j - n_{1j})^{-1}\sum_{i=1}^{n} I(\widehat{e}_i \in \widehat{Q}_j)\{(1-Z_i)Y_i - \overline{y}_{0j}\}^2$, $\overline{y}_{1j} = n_{1j}^{-1}\sum_{i=1}^{n} I(\widehat{e}_i \in \widehat{Q}_j)Z_iY_i$, and $\overline{y}_{0j} = (n_j - n_{1j})^{-1}\sum_{i=1}^{n} I(\widehat{e}_i \in \widehat{Q}_j)(1-Z_i)Y_i$. Use of (4.1) is equivalent to assuming one can obtain independent samples from the true strata of e, and $(Y_0, Y_1) \perp\!\!\!\perp Z$ holds within strata.

Table 4.2 and Table 4.3 display Monte Carlo ARE results for several sample sizes along with absolute percent bias, Monte Carlo standard error, and average estimated standard error. Points indicating Monte Carlo ARE of $\widehat{\Delta}_{IPW1}$ and $\widehat{\Delta}_{IPW2}$ relative to $\widehat{\Delta}_S$ are displayed in relation to their large sample predictions in Figures 4.1 and 4.2. Clearly the asymptotic predictions do not reflect performance in smaller samples when X has strong associations with treatment and prognosis. We see poor performance of $\widehat{\Delta}_S$ at small sample sizes, especially compared to $\widehat{\Delta}_{IPW2}$. Although the variances of $\widehat{\Delta}_S$, $\widehat{\Delta}_{IPW2}$, and $\widehat{\Delta}_{DR}$ are comparable, the small sample biases of $\widehat{\Delta}_S$ for our simple model are substantially larger than the asymptotic bias. For the stratification estimator, larger sample sizes are needed to achieve reasonable covariate balance within strata.

As expected, $\widehat{\Delta}_{DR}$ is the most efficient of the weighted estimators relative to $\widehat{\Delta}_S$. If we exclude $\widehat{\Delta}_{DR}$, and consider the two weighted estimators that only model $e(X)$,

then $\widehat{\Delta}_{IPW2}$ appears preferable to $\widehat{\Delta}_{IPW1}$ on grounds of smaller variance. Heuristically, the use of $(\sum_{i=1}^{n} Z_i/\widehat{e}_i)^{-1}$ and $\{\sum_{i=1}^{n}(1-Z_i)/(1-\widehat{e}_i)\}^{-1}$ in $\widehat{\Delta}_{IPW2}$ to average the $Z_iY_i/\widehat{e}_i$ and $\{(1-Z_i)Y_i\}/(1-\widehat{e}_i)$, as opposed to the usual n^{-1} used by $\widehat{\Delta}_{IPW1}$, functions to wash out the effects of large weights, giving $\widehat{\Delta}_{IPW2}$ greater stability overall. Indeed, if μ_1 and μ_0 are assumed to have the same sign, the difference between (3.1) and (3.2) is positive except in some cases of extreme values for either $\mathrm{Cov}\{E(Y_1|X), e(X)^{-1}\}$ or $\mathrm{Cov}[E(Y_0|X), \{1-e(X)\}^{-1}]$.

Monte Carlo estimates of coverage for the various estimators are displayed in Table 4.4 for both binary and continuous responses. Coverage rates decrease with increasing values of α_1 or β_1 and, with the exception of $\widehat{\Delta}_S$, approach the nominal level as sample size increases. The contents of Tables 4.2 and 4.3 show the average of the standard errors for $\widehat{\Delta}_S$ computed using (4.1) closely approximate the Monte Carlo values as n becomes large, and hence the low coverage percentages shown in Table 4.4 are due to the residual bias of $\widehat{\Delta}_S$. It also appears that average standard errors for both $\widehat{\Delta}_S$ and $\widehat{\Delta}_{IPW1}$ exhibit a slightly stronger tendency to underestimate variability at smaller sample sizes when compared to $\widehat{\Delta}_{IPW2}$ and $\widehat{\Delta}_{DR}$ for the parameter settings examined here.

The purpose of stratification is to produce treatment groups more comparable in their distribution of the confounding covariates. Larger sample sizes allow one to refine this effect through an increase in the number of strata used. Table 4.5 shows how performance of $\widehat{\Delta}_S$ changes when K is increased from five to ten at the sample size of 5000. These $K = 10$ results can be compared to the $K = 5$ results of Tables

4.2 and 4.3. For both binary and continuous responses the absolute percent bias of $\widehat{\Delta}_S$ was reduced by approximately sixty percent at each of the parameter settings.

Coverage percentages for $\widehat{\Delta}_S$ also improved dramatically at $K = 10$. Note that Monte Carlo standard errors and their estimates remained fairly constant from $K = 5$ to $K = 10$, indicating this improvement in coverage comes solely from the reduction in bias achieved by using more strata. Although not presented here, an interesting problem would be to determine the rate at which K should increase with n to minimize the potential bias inherent in $\widehat{\Delta}_S$.

4.2 Estimation of $\boldsymbol{\beta}$

Here we present an empirical example illustrating the effect of estimating versus knowing the true value of the parameter $\boldsymbol{\beta}$ in the propensity model as discussed in Section 3.2. In this case the confounder $\boldsymbol{X} = (X_1, X_2)^T$ is bivariate normal with mean and variance parameters

$$\mu_X = \begin{pmatrix} 2 \\ 2 \end{pmatrix} \qquad \Sigma_X = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

The propensity model is $e(\boldsymbol{X}) = \{1+\exp(-\beta_0-\beta_1 X_1-\beta_2 X_2)\}^{-1}$. For continuous Y, the response conditional on $\boldsymbol{X}$ and Z is $N(\alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \alpha_3 Z, 1)$. For binary Y, the conditional mean is $\{1 + \exp(-\alpha_0 - \alpha_1 X_1 - \alpha_2 X_2 - \alpha_3 Z)\}^{-1}$. The values of α_0, β_0, α_1, β_1, and α_3 are fixed at 0, 0, 0.5, 0.5, and 1, respectively.

The empirical results presented in Table 4.6 verify the reduction in variance predicted for the weighted estimators through estimation of $\boldsymbol{\beta}$ and the absence of any

change in variability for $\widehat{\Delta}_{DR}$ as discussed in Section 3.1. The results for $\widehat{\Delta}_S$ suggest that $\Gamma_{\beta qp}$, the piece of Σ_S due to estimating $\boldsymbol{\beta}$, tends to be negative. We have evaluated $\Gamma_{\beta qp}$ for several parameter settings under the above model and found the sign to be negative in all cases.

The gains when estimating $\boldsymbol{\beta}$ presented in Table 4.6 are not always as dramatic for $\widehat{\Delta}_S$ as those for $\widehat{\Delta}_{IPW1}$ and $\widehat{\Delta}_{IPW2}$. For $\widehat{\Delta}_S$, the value of $\boldsymbol{\beta}$ is only used to determine strata boundaries. Thus, heuristically, estimating or not estimating its value would not be expected to influence stability as dramatically as it does for the IPW estimators, where the $\widehat{e}_i$ are used to weight each observation.

4.3 Extra Covariates in the Propensity Score

To illustrate the effect of adding an additional covariate to the propensity model when that covariate is not associated with treatment exposure given the covariates already present, the same models and covariates detailed in the previous section can be used. In this case, to be consistent with the notation used in Section 3.3, let $X = X_1$, $V = X_2$, and $\gamma = \beta_2$. Here the values of α_0, β_0, α_1, β_1, and α_3 are fixed at 0, 0, 0.5, 1, and 1, respectively. Because we desire $P(Z = 1|X, V) = P(Z = 1|X)$, the value of γ is set at zero. With these settings, performance of the various estimators when V is or is not incorporated into e is displayed in Table 4.7. Results are shown for varying levels of association between V and the response as specified through the value of α_2.

A technical consideration to note here is that when comparing $\widehat{\Delta}_{DR}$ with or without inclusion of V one must be using compatible $E(Y|X,V,Z)$ and $E(Y|X,Z)$. In the case of the normally distributed response, integrating $E(Y|X,V,Z)$ over the distribution of V yields another linear model. However, in the binary case if one starts with the logistic regression model for $E(Y|X,V,Z)$, no closed form expression for $E(Y|X,Z)$ with parameters to be estimated is available. However, our primary interest is comparison of $\widehat{\Delta}_S$ to $\widehat{\Delta}_{IPW1}$ and $\widehat{\Delta}_{IPW2}$; so we avoid this technicality and omit consideration of $\widehat{\Delta}_{DR}$ for the "V not included" portions of Table 4.7.

The results in Table 4.7 confirm the reduction in variance expected for the IPW estimators when "over fitting" the propensity score using prognostic covariates. It appears that $\widehat{\Delta}_S$ follows this same pattern. We have evaluated $\Gamma_{\gamma\beta qp}$ for several parameter settings and found the sign to be negative in all cases where V is associated with Y.

4.4 Discussion

The large sample bias of $\widehat{\Delta}_S$ will always dominate the ARE ratio when sample size increases and the number of strata remains fixed. Using more strata will increase the sample sizes at which the trade-off in bias and variance takes place, but stratifying on quintiles seems to be the most popular approach in practice—even for substantial sample sizes. Arguments for stratification are that it reduces variability by avoiding the large weights of extreme $\widehat{e}_i$ and is an approach that perhaps does not rely as heavily

on correct specification of the propensity model because predictions are only used to determine strata cut-off points. There is also the potential interest in estimating within strata treatment differences and the practical utility of checking covariate balance within strata as a propensity model development strategy. However, the presentation here would support reporting weighted estimators of Δ in most cases where the stratification estimator is reported, with use of $\widehat{\Delta}_{IPW2}$ preferable to $\widehat{\Delta}_{IPW1}$.

4.5 Figures and Tables

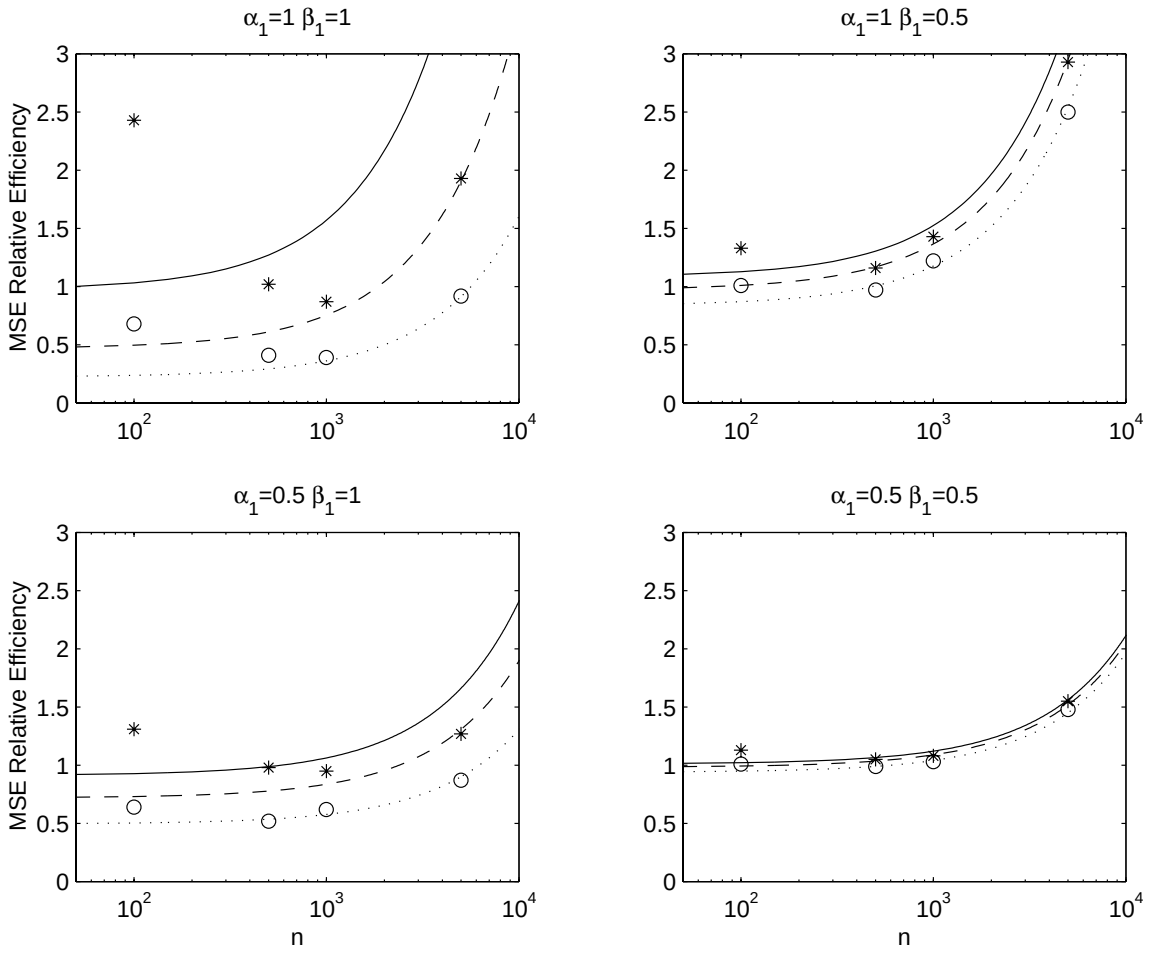

Figure 4.1: ARE profiles for a continuous response. Dotted lines are ARE_{IPW1}. Dashed lines are ARE_{IPW2}. Solid lines are ARE_{DR}. Monte Carlo results are shown for sample sizes 100, 500, 1000, and 5000, where open circles indicate Monte Carlo ARE_{IPW1} values and asterisks indicate Monte Carlo ARE_{IPW2} values.

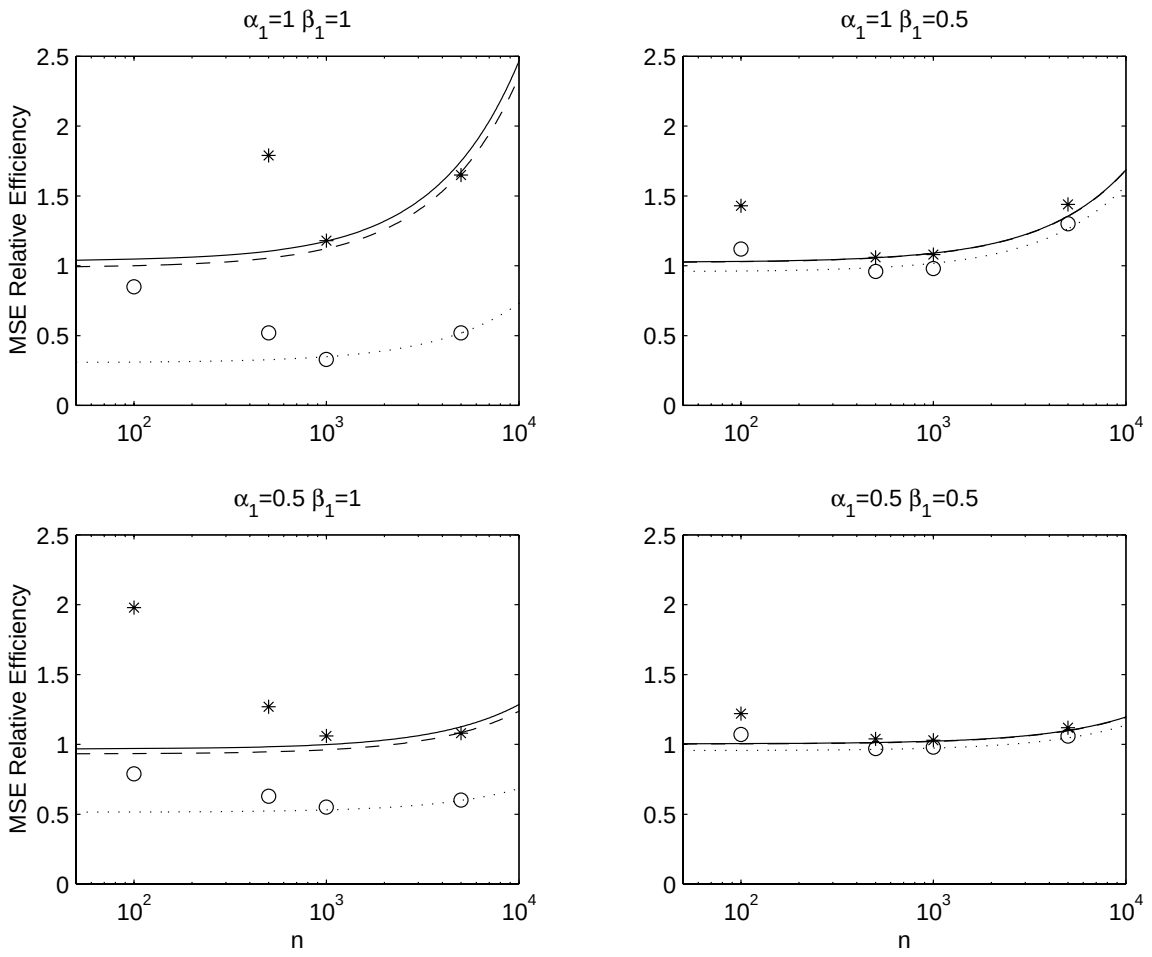

Figure 4.2: ARE profiles for a binary response. Dotted lines are ARE_{IPW1}. Dashed lines are ARE_{IPW2}. Solid lines are ARE_{DR}. Monte Carlo results are shown for sample sizes 100, 500, 1000, and 5000, where open circles indicate Monte Carlo ARE_{IPW1} values and asterisks indicate Monte Carlo ARE_{IPW2} values.

Table 4.1: Asymptotic results for a univariate confounder.

α_1	β_1	Δ	Bias$(\widehat{\Delta}_S)$	Σ_S	Σ_{IPW1}	Σ_{IPW2}	Σ_{DR}
			continuous Y				
1.0	1.0	1	0.093	14.01	62.48	29.97	14.41
	0.5	1	0.050	5.96	7.12	6.14	5.50
0.5	1.0	1	0.046	13.2	26.5	18.3	14.41
	0.5	1	0.025	5.56	5.92	5.66	5.50
			binary Y				
1.0	1.0	0.086	0.012	0.96	3.15	0.98	0.93
	0.5	0.086	0.006	0.49	0.51	0.48	0.48
0.5	1.0	0.150	0.008	2.10	4.08	2.25	2.17
	0.5	0.150	0.004	0.89	0.93	0.89	0.89

Table 4.2: Monte Carlo ARE results for a univariate confounder and a continuous response. Bias is the absolute percent bias of the relevant estimator. Numbers in parentheses are average estimated standard errors.

		Bias				Standard Error				ARE		
α_1	β_1	$\widehat{\Delta}_S$	$\widehat{\Delta}_{IPW1}$	$\widehat{\Delta}_{IPW2}$	$\widehat{\Delta}_{DR}$	$\widehat{\Delta}_S$	$\widehat{\Delta}_{IPW1}$	$\widehat{\Delta}_{IPW2}$	$\widehat{\Delta}_{DR}$	$IPW1$	$IPW2$	DR
						$n = 500$						
1.0	1.0	10.5	1.6	1.8	0.6	0.212 (0.146)	0.370 (0.283)	0.234 (0.198)	0.174	0.41	1.02	1.86
	0.5	4.9	0.1	0.1	0.1	0.112 (0.108)	0.124 (0.118)	0.113 (0.110)	0.106	0.97	1.16	1.33
0.5	1.0	5.5	0.3	0.7	0.3	0.183 (0.142)	0.265 (0.201)	0.193 (0.168)	0.178	0.52	0.98	1.14
	0.5	2.4	0.1	0.1	0.0	0.108 (0.105)	0.111 (0.108)	0.108 (0.106)	0.106	0.99	1.05	1.08
						$n = 1000$						
1.0	1.0	9.5	1.2	1.3	0.2	0.126 (0.114)	0.253 (0.216)	0.169 (0.150)	0.121	0.39	0.87	1.71
	0.5	5.0	0.2	0.2	0.1	0.078 (0.077)	0.084 (0.084)	0.077 (0.078)	0.074	1.22	1.43	1.58
0.5	1.0	4.8	0.6	0.7	0.2	0.121 (0.111)	0.165 (0.147)	0.133 (0.124)	0.121	0.63	0.95	1.16
	0.5	2.4	0.0	0.1	0.1	0.075 (0.074)	0.077 (0.077)	0.075 (0.075)	0.074	1.02	1.08	1.12
						$n = 5000$						
1.0	1.0	9.4	0.2	0.3	0.1	0.053 (0.053)	0.112 (0.106)	0.077 (0.074)	0.054	0.92	1.94	4.00
	0.5	4.9	0.0	0.0	0.0	0.035 (0.034)	0.038 (0.038)	0.035 (0.035)	0.033	2.51	2.93	3.32
0.5	1.0	4.5	0.2	0.1	0.2	0.052 (0.051)	0.074 (0.071)	0.061 (0.059)	0.055	0.87	1.27	1.60
	0.5	2.5	0.0	0.0	0.0	0.033 (0.033)	0.034 (0.034)	0.033 (0.034)	0.033	1.49	1.55	1.59

Table 4.3: Monte Carlo ARE results for a univariate confounder and binary response. Bias is the absolute percent bias of the relevant estimator. Numbers in parentheses are average estimated standard errors.

		Bias				Standard Error				ARE		
α_1	β_1	$\widehat{\Delta}_S$	$\widehat{\Delta}_{IPW1}$	$\widehat{\Delta}_{IPW2}$	$\widehat{\Delta}_{DR}$	$\widehat{\Delta}_S$	$\widehat{\Delta}_{IPW1}$	$\widehat{\Delta}_{IPW2}$	$\widehat{\Delta}_{DR}$	$IPW1$	$IPW2$	DR
						$n = 500$						
1.0	1.0	20.5	7.0	2.1	0.5	0.058 (0.040)	0.083 (0.069)	0.045 (0.042)	0.044	0.53	1.79	1.90
	0.5	5.7	0.3	0.7	0.7	0.031 (0.031)	0.032 (0.032)	0.031 (0.031)	0.031	0.96	1.06	1.06
0.5	1.0	9.3	1.8	1.3	0.4	0.075 (0.058)	0.096 (0.082)	0.067 (0.063)	0.066	0.63	1.27	1.30
	0.5	2.6	0.1	0.1	0.2	0.043 (0.042)	0.044 (0.043)	0.042 (0.042)	0.042	0.97	1.03	1.04
						$n = 1000$						
1.0	1.0	13.6	1.0	0.8	0.2	0.033 (0.030)	0.061 (0.052)	0.032 (0.030)	0.031	0.33	1.18	1.25
	0.5	6.0	0.5	0.5	0.5	0.022 (0.022)	0.023 (0.023)	0.022 (0.022)	0.022	0.99	1.08	1.08
0.5	1.0	5.4	0.5	0.2	0.3	0.048 (0.044)	0.066 (0.060)	0.047 (0.046)	0.047	0.55	1.06	1.10
	0.5	2.8	0.3	0.1	0.1	0.030 (0.030)	0.031 (0.030)	0.030 (0.030)	0.030	0.97	1.03	1.04
						$n = 5000$						
1.0	1.0	13.4	0.3	0.0	0.1	0.014 (0.014)	0.025 (0.025)	0.014 (0.014)	0.014	0.52	1.68	1.75
	0.5	6.4	0.0	0.1	0.1	0.010 (0.010)	0.010 (0.010)	0.010 (0.010)	0.010	1.26	1.34	1.35
0.5	1.0	5.7	0.4	0.3	0.2	0.021 (0.020)	0.029 (0.028)	0.021 (0.021)	0.021	0.60	1.09	1.12
	0.5	2.8	0.1	0.1	0.1	0.013 (0.013)	0.013 (0.014)	0.013 (0.013)	0.013	1.06	1.11	1.11

Table 4.4: Coverage percentages for continuous and binary responses.

		Y continuous			Y binary		
α_1	β_1	$\widehat{\Delta}_S$	$\widehat{\Delta}_{IPW1}$	$\widehat{\Delta}_{IPW2}$	$\widehat{\Delta}_S$	$\widehat{\Delta}_{IPW1}$	$\widehat{\Delta}_{IPW2}$
				$n = 500$			
1.0	1.0	80.0	87.9	89.4	89.3	90.1	91.9
	0.5	91.9	93.1	94.2	95.0	93.7	95.1
0.5	1.0	84.4	90.1	91.9	86.2	91.1	92.4
	0.5	93.5	94.2	94.4	94.2	94.3	94.5
				$n = 1000$			
1.0	1.0	82.6	89.8	90.9	93.6	92.3	93.1
	0.5	89.4	94.8	95.1	94.4	93.7	94.5
0.5	1.0	89.3	92.1	92.9	92.1	92.7	93.2
	0.5	93.9	94.7	95.1	94.6	94.6	94.8
				$n = 5000$			
1.0	1.0	57.1	92.7	93.0	88.7	94.4	94.6
	0.5	70.3	94.6	94.8	91.9	95.0	95.0
0.5	1.0	84.2	93.9	94.1	93.1	94.5	94.6
	0.5	88.2	95.0	95.0	94.2	95.7	95.3

Table 4.5: Effect of increasing K from five to ten. Bias is the absolute percent bias of $\widehat{\Delta}_S$. Numbers in parentheses are average estimated standard errors. Sample size is 5000.

					ARE		
α_1	β_1	Bias	Standard Error	Coverage	$IPW1$	$IPW2$	DR
				continuous Y			
1.0	1.0	3.7	0.054 (0.053)	88.3	0.34	0.73	1.50
	0.5	2.0	0.034 (0.034)	90.8	1.06	1.24	1.41
0.5	1.0	1.7	0.054 (0.052)	92.6	0.49	0.86	1.09
	0.5	1.0	0.033 (0.033)	94.1	1.02	1.07	1.10
				binary Y			
1.0	1.0	5.2	0.014 (0.014)	94.6	0.34	1.10	1.15
	0.5	2.6	0.010 (0.010)	94.3	0.99	1.06	1.06
0.5	1.0	2.3	0.021 (0.021)	94.2	0.56	1.01	1.04
	0.5	1.1	0.013 (0.013)	95.3	0.98	1.02	1.02

Table 4.6: Monte Carlo ARE results for estimating $\boldsymbol{\beta}$ for continuous and binary responses. Bias is the absolute percent bias of the relevant estimator. Numbers in parentheses are results observed when $\boldsymbol{\beta}$ is known. Sample size is 2500.

		Bias				Standard Error				ARE		
α_2	β_2	$\widehat{\Delta}_S$	$\widehat{\Delta}_{IPW1}$	$\widehat{\Delta}_{IPW2}$	$\widehat{\Delta}_{DR}$	$\widehat{\Delta}_S$	$\widehat{\Delta}_{IPW1}$	$\widehat{\Delta}_{IPW2}$	$\widehat{\Delta}_{DR}$	$IPW1$	$IPW2$	DR
						continuous Y						
0.5	0.5	4.9	0.2	0.1	0.0	0.069	0.097	0.077	0.068	0.75	1.19	1.55
		(4.9)	(0.3)	(0.1)	(0.0)	(0.069)	(0.190)	(0.089)	(0.067)	(0.20)	(0.90)	(1.57)
	-0.5	0.1	0.1	0.1	0.1	0.043	0.047	0.043	0.043	0.83	0.97	0.98
		(0.0)	(0.1)	(0.0)	(0.1)	(0.052)	(0.120)	(0.053)	(0.043)	(0.19)	(0.97)	(1.46)
-0.5	0.5	0.0	0.1	0.1	0.2	0.072	0.074	0.074	0.070	0.94	0.95	1.05
		(0.0)	(0.0)	(0.0)	(0.2)	(0.083)	(0.085)	(0.085)	(0.069)	(0.95)	(0.96)	(1.45)
	-0.5	4.7	0.0	0.0	0.0	0.043	0.043	0.043	0.042	2.17	2.16	2.25
		(4.7)	(0.0)	(0.0)	(0.0)	(0.043)	(0.54)	(0.054)	(0.042)	(1.38)	(1.39)	(2.24)
						binary Y						
0.5	0.5	7.5	0.4	0.4	0.3	0.020	0.025	0.019	0.019	0.65	1.11	1.12
		(7.5)	(0.6)	(0.6)	(0.3)	(0.020)	(0.065)	(0.020)	(0.019)	(0.10)	(1.06)	(1.13)
	-0.5	0.1	0.0	0.1	0.1	0.012	0.014	0.012	0.012	0.77	0.97	0.98
		(0.0)	(1.0)	(0.1)	(0.2)	(0.013)	(0.041)	(0.013)	(0.012)	(0.09)	(0.97)	(1.02)
-0.5	0.5	0.1	0.0	0.1	0.1	0.032	0.034	0.033	0.032	0.91	0.97	0.99
		(0.2)	(0.5)	(0.2)	(0.1)	(0.033)	(0.049)	(0.034)	(0.032)	(0.46)	(0.97)	(1.06)
	-0.5	4.5	0.2	0.2	0.2	0.20	0.20	0.20	0.20	1.18	1.22	1.23
		(4.4)	(0.2)	(0.2)	(0.2)	(0.020)	(0.33)	(0.021)	(0.020)	(0.44)	(1.08)	(1.23)

Table 4.7: Monte Carlo results for incorporating an extra covariate into the propensity score for continuous and binary responses. Bias is the absolute percent bias of the relevant estimator. Numbers in parentheses are results observed when extra covariate is not included in the propensity model. Sample size is 2500.

	Bias				Standard Error				ARE		
α_2	$\widehat{\Delta}_S$	$\widehat{\Delta}_{IPW1}$	$\widehat{\Delta}_{IPW2}$	$\widehat{\Delta}_{DR}$	$\widehat{\Delta}_S$	$\widehat{\Delta}_{IPW1}$	$\widehat{\Delta}_{IPW2}$	$\widehat{\Delta}_{DR}$	$IPW1$	$IPW2$	DR
						continuous Y					
-1.0	4.8	0.2	0.4	0.1	0.087	0.092	0.098	0.077	1.17	1.03	1.68
	(4.9)	(0.2)	(0.4)		(0.105)	(0.110)	(0.115)		(1.11)	(1.02)	
0.0	4.6	0.0	0.1	0.0	0.073	0.103	0.085	0.076	0.70	1.04	1.31
	(4.6)	(0.0)	(0.1)		(0.073)	(0.103)	(0.084)		(0.71)	(1.05)	
1.0	4.8	0.6	0.5	0.3	0.088	0.161	0.100	0.078	0.39	1.00	1.65
	(4.8)	(0.7)	(0.5)		(0.106)	(0.170)	(0.116)		(0.47)	(1.00)	
						binary Y					
-1.0	4.5	0.5	0.6	0.3	0.033	0.039	0.036	0.033	0.77	0.93	1.05
	(4.4)	(0.5)	(0.5)		(0.035)	(0.040)	(0.037)		(0.78)	(0.93)	
0.0	5.3	0.1	0.1	0.2	0.029	0.042	0.030	0.030	0.54	1.01	1.05
	(5.3)	(0.1)	(0.1)		(0.029)	(0.041)	(0.030)		(0.53)	(1.00)	
1.0	6.5	1.8	0.2	0.1	0.016	0.034	0.016	0.016	0.23	1.01	1.03
	(6.7)	(2.1)	(0.4)		(0.016)	(0.034)	(0.016)		(0.24)	(1.02)	

Chapter 5

Two-stage Randomization Designs in Clinical Trials

This chapter introduces the underlying model framework of the approach to estimating survival in two-stage randomized trials and several proposed estimators of survival distribution and mean survival time in the presence of right censoring. Arguments for the large sample properties of the estimators are given in Appendix B. In Chapter 6, the proposed methods are applied to the CALGB 8923 data and performance of the estimators is investigated via simulation.

5.1 Model Framework and Proposed Estimators

Let treatment A, at levels A_1 and A_2, and treatment B, at levels B_1 and B_2, be the induction and post-remission treatments, respectively. In the trial, patients are

randomized initially to one of the A treatment levels. If a patient achieves remission and consents to further participation in the trial, s/he is then randomized to a level of B. Consider four continuous survival-time random variables: $T_{11i}, T_{12i}, T_{21i}$, and T_{22i}, where $T_{jki},\ j,k = 1,2$, represents survival time for the ith individual under policy "treat with A_j followed by B_k if remission and consent," which we henceforth write as "A_jB_k." These variables are the potential survival outcomes under the treatment policies to which individual i might have been exposed, possibly contrary to that actually received by that individual in the trial. Thus, the full set of survival times (T_{11}, T_{12}, T_{21},T_{22}) is not observable for each individual in the trial, and the T_{jk} may be viewed as counterfactuals (e.g. Holland, 1986).

This counterfactual representation provides a framework in which comparisons among treatment policies may be characterized clearly. In particular, the distributions of the T_{jk} represent survival in the population were all patients to be assigned to policy A_jB_k, so that inference on their features addresses directly the "intent-to-treat" question of interest. Thus, our approach is to estimate the survival distributions and means for the T_{jk} on the basis of the observed data from a two-stage randomization trial. It is critical to recognize that the use of counterfactuals here does not imply a focus on causal treatment effects (Holland 1986); rather, we use counterfactuals as a convenient device with which to state and address the "intent-to-treat" question.

Because patients are randomized to A, the data used to estimate quantities related to T_{11} and T_{12} are independent of those used for estimating characteristics of T_{21} and T_{22}. Thus, for simplicity, we restrict attention to the A_1 data here and in Appendix

B; the development would be identical for A_2. Accordingly, $i = 1, \ldots, n$ in the sequel refers to the n patients randomized to A_1; e.g. for 50-50 A randomization, the total number of patients is roughly $2n$. If there were no censoring, the observable random variables for the ith patient would be T_i, survival time under the treatment policy actually assigned; R_i, a remission/consent status indicator ($R_i = 0$ if no remission or consent, $R_i = 1$ if remission and consent); and, if $R_i = 1$, Z_i, the B treatment assignment indicator ($Z_i = 0$ if assigned to B_1, Z_i=1 if assigned to B_2). We make the reasonable assumption that subjects whose disease does not remit or who do not consent would have equal potential responses under either A_1B_1 or A_1B_2; i.e. when $R_i = 0$, $T_{11i} = T_{12i}$. Under this scheme, we assume T_i satisfies the relationship

$$\begin{aligned} T_i &= (1 - R_i)T_{11i} + R_i(1 - Z_i)T_{11i} + R_iZ_iT_{12i} \\ &= (1 - R_i)T_{12i} + R_i(1 - Z_i)T_{11i} + R_iZ_iT_{12i}. \end{aligned}$$

Thus, in the case of no censoring, the observable data would be the i.i.d. vectors (R_i, R_iZ_i, T_i).

To account for right censoring, let C_i be the time to censoring and let $K(t) = P(C_i > t)$. Then the full, though unobservable, data are the i.i.d. vectors $(R_i, R_iZ_i, C_i, T_{11i}, T_{12i})$. Because for most clinical trials total follow-up of patients is limited, of necessity we can only consider restricted lifetime; that is, survival up to some time L, where L is smaller than the maximum follow-up time. Restricted lifetime is defined formally as $T_{jk}^L = \min(T_{jk}, L)$, where $K(L) > 0$. For simplicity, we denote T_{jk}^L as T_{jk}, and this convention should be kept in mind. Under these condi-

tions, then, we observe i.i.d. vectors $(R_i, R_i Z_i, V_i, \Delta_i)$, where $\Delta_i = I(T_i < C_i)$ is the censoring indicator; $V_i = \min(T_i, C_i)$ is the observed time until death or censoring; and, if an individual is censored but has not yet remitted, we take $R_i = 0$. Assume C_i is independent of $(R_i, R_i Z_i, T_{11i}, T_{12i})$ and, conditional on $R_i = 1$, Z_i is independent of T_{11i} and T_{12i}, which is trivially satisfied by randomization of B treatments.

Consider estimation of $S_{1k}(t) = 1 - P(T_{1k} \leq t) = 1 - F_{1k}(t)$, the survival function under policy $A_1 B_k$ for $k = 1, 2$. Note that $S_{1k}(t)$ is the same whether we consider actual or restricted lifetime as long as $t < L$. Let $\pi_Z = P(Z_i = 1 | R_i = 1)$, which we take to be known, and write $Q_{1i} = 1 - R_i + (1 - \pi_Z)^{-1} R_i (1 - Z_i)$ and $Q_{2i} = 1 - R_i + \pi_Z^{-1} R_i Z_i$. Then, for $k = 1$,

$$\begin{aligned}
E\left\{\frac{\Delta_i Q_{1i}}{K(V_i)} I(V_i \leq t)\right\} &= E\left\{\frac{I(T_{11i} < C_i) Q_{1i}}{K(T_{11i})} I(T_{11i} \leq t)\right\} \\
&= E\left[\frac{I(T_{11i} \leq t)}{K(T_{11i})} Q_{1i} E\{I(T_{11i} \leq C_i) | R_i, R_i Z_i, T_{11i}\}\right] \\
&= E\left\{\frac{I(T_{11i} \leq t)}{K(T_{11i})} K(T_{11i}) Q_{1i}\right\} = E\{I(T_{11i} \leq t) Q_{1i}\} \\
&= E[E\{I(T_{11i} \leq t) Q_{1i} | R_i, T_{11i}\}] \\
&= E\{I(T_{11i} \leq t) E(Q_{1i} | R_i, T_{11i})\} \\
&= F_{11}(t),
\end{aligned}$$

which follows by noting that

$$E(Q_{1i} | R_i, T_{11i}) = 1 - R_i + (1 - \pi_Z)^{-1} E\{R_i (1 - Z_i) | R_i, T_{11i}\} = 1$$

from considering the cases $R_i = 0$ and $R_i = 1$ in turn. The argument for $k = 2$ is identical.

These developments immediately suggest the estimator

$$\widehat{F}_{1k}(t) = n^{-1} \sum_{i=1}^{n} \frac{\Delta_i Q_{ki}}{\widehat{K}(V_i)} I(V_i \leq t), \tag{5.1}$$

where $\widehat{K}(t) = \prod_{u \leq t} \{1 - dN^c(u)/Y(u)\}$, is the Kaplan-Meier estimate of the censoring survivor curve, with $N^c(u) = \sum_{i=1}^{n} I(V_i \leq u, \Delta_i = 0)$ and $Y(u) = \sum_{i=1}^{n} I(V_i \geq u)$. The estimator $\widehat{F}_{1k}(t)$ thus incorporates two forms of weighting of "complete cases" (e.g. Robins, Rotnitzky, and Zhao, 1994). Probabilistically, every uncensored value of T_i represents $K(T_i)^{-1}$ individuals, so that, roughly speaking, the response for an uncensored individual counts for him/herself and $\{K(T_i)^{-1} - 1\}$ similar, censored individuals. Likewise, Q_{ki} involves differential weighting of remitting/consenting and other patients. A patient who achieves remission and consents and is randomized to one of the B treatments with probability π_Z represents π_Z^{-1} individuals who potentially could have been assigned that B treatment.

A second estimator may be deduced following the observation that inverse weighting of uncensored cases duplicates such cases according to the inverse probability of observation. This suggests, instead of the usual average based on n individuals, averaging using a probabilistically-adjusted sample size; i.e.

$$\widehat{F}'_{1k}(t) = \left\{ \sum_{i=1}^{n} \frac{\Delta_i Q_{ki}}{\widehat{K}(V_i)} \right\}^{-1} \sum_{i=1}^{n} \frac{\Delta_i Q_{ki}}{\widehat{K}(V_i)} I(V_i \leq t). \tag{5.2}$$

Because if there are no ties and the largest observation is a failure, which should be true with high probability under the assumptions above, $\sum_{i=1}^{n} \Delta_i / \widehat{K}(V_i) = n$, (5.1)

may be written alternatively as the solution to the estimating equation

$$\sum_{i=1}^{n} \frac{\Delta_i}{\widehat{K}(V_i)} \{Q_{ki} I(V_i \leq t) - F_{1k}(t)\} = 0. \tag{5.3}$$

Similarly, (5.2) may be written as the solution to

$$\sum_{i=1}^{n} \frac{\Delta_i}{\widehat{K}(V_i)} [Q_{ki} \{I(V_i \leq t) - F_{1k}(t)\}] = 0. \tag{5.4}$$

Combining (5.3) and (5.4), Both estimators may be written as solutions to equations of the form

$$\sum_{i=1}^{n} \frac{\Delta_i}{\widehat{K}(V_i)} \{Q_{ki} I(V_i \leq t) - F_{1k}(t) - \alpha_{1k}(Q_{ki} - 1)\} = 0, \tag{5.5}$$

where $\alpha_{1k} = 0$ yields the equation for $\widehat{F}_{1k}(t)$ and $\alpha_{1k} = F_{1k}(t)$ gives that for $\widehat{F}'_{1k}(t)$. This suggests a third estimator, $\widehat{F}''_{1k}(t)$, say, where α_{1k} is chosen to have minimum variance among all estimators solving equations of the form (5.5). It is shown in Appendix B that

$$\widehat{F}''_{1k}(t) = n^{-1} \sum_{i=1}^{n} \frac{\Delta_i Q_{ki}}{\widehat{K}(V_i)} I(V_i \leq t) - \widehat{\alpha}_{1k}\, n^{-1} \sum_{i=1}^{n} \frac{\Delta_i}{\widehat{K}(V_i)} (Q_{ki} - 1), \tag{5.6}$$

where $\widehat{\alpha}_{1k}$ equals

$$\frac{n^{-1} \sum_{i=1}^{n} \Delta_i Q_{ki}(Q_{ki} - 1) I(V_i \leq t)/\widehat{K}(V_i) + \int_0^L dN^c(u) \{\widehat{K}(u) Y(u)\}^{-1} \widehat{E}\, \{L^{\alpha}_{1k}(t,u)\}}{n^{-1} \sum_{i=1}^{n} (Q_{ki} - 1)^2 + \int_0^L dN^c(u) \{\widehat{K}(u) Y(u)\}^{-1} \widehat{E}\, \{G^{\alpha}_{k}(u)\}}, \tag{5.7}$$

$\widehat{E}\{L^{\alpha}_{1k}(t,u)\} = n^{-1} \sum_{i=1}^{n} \Delta_i \{Q_{ki} I(V_i \leq t) - \widehat{G}_{1k}(t,u)\}\{Q_{ki} - 1 - \widehat{G}_{Q_k}(u)\} I(V_i \geq u)/\widehat{K}(V_i)$, $\widehat{E}\, \{G^{\alpha}_{k}(u)\} = n^{-1} \sum_{i=1}^{n} \Delta_i \{Q_{ki} - 1 - \widehat{G}_{Q_k}(u)\}^2 I(V_i \geq u)/\widehat{K}(V_i)$, $\widehat{G}_{Q_k}(u) = \{n\widehat{S}(u)\}^{-1} \times \sum_{i=1}^{n} \Delta_i (Q_{ki} - 1) I(V_i \geq u)/\widehat{K}(V_i)$, $\widehat{G}_{1k}(t,u) = \{n\widehat{S}(u)\}^{-1}$

$\times \sum_{i=1}^{n} \Delta_i Q_{ki} I(V_i \leq t) I(V_i \geq u)/\widehat{K}(V_i)$, and $\widehat{S}(u)$ is the Kaplan-Meier estimator for $P(T > u)$.

Substituting V_i for $I(V_i \leq t)$ in (5.1), (5.2), and (5.6) leads to analogous estimators for mean (restricted) survival time, which we denote as $\widehat{\mu}_{1k}$, $\widehat{\mu}'_{1k}$, and $\widehat{\mu}''_{1k}$, respectively.

Arguments for consistency and asymptotic normality of the estimators are given in Appendix B and yield variance estimators for (5.1), (5.2), and (5.6), leading to variance estimators for $\widehat{S}_{1k}(t) = 1 - \widehat{F}_{1k}(t)$, $\widehat{S}'_{1k}(t) = 1 - \widehat{F}'_{1k}(t)$, and $\widehat{S}''_{1k}(t) = 1 - \widehat{F}''_{1k}(t)$ given by

$$\widehat{\text{var}}\{\widehat{S}_{1k}(t)\}$$
$$= \quad n^{-1}\left[n^{-1}\sum_{i=1}^{n}\frac{\Delta_i Q_{ki}}{\widehat{K}(V_i)} I(V_i \leq t) - \widehat{F}_{1k}(t)^2 + \int_0^L \frac{dN^c(u)}{\widehat{K}(u)Y(u)}\widehat{E}\{L_{1ki}(t,u)\}^2\right], \quad (5.8)$$

$$\widehat{\text{var}}\{\widehat{S}'_{1k}(t)\}$$
$$= \quad n^{-1}\left[n^{-1}\sum_{i=1}^{n}\frac{\Delta_i Q_{ki}}{\widehat{K}(V_i)}\{I(V_i \leq t) - \widehat{F}'_{1k}(t)\}^2 + \int_0^L \frac{dN^c(u)}{\widehat{K}(u)Y(u)}\widehat{E}\{L'_{1ki}(t,u)\}^2\right], \quad (5.9)$$

and

$$\widehat{\text{var}}\{\widehat{S}''_{1k}(t)\} \quad = \quad n^{-1}\left[n^{-1}\sum_{i-1}^{n}\frac{\Delta_i}{\widehat{K}(V_i)}\{Q_{ki}I(V_i \leq t) - \widehat{F}''_{1k}(t) - \widehat{\alpha}_{1k}(Q_{ki} - 1)\}^2 + \int_0^L \frac{dN^c(u)}{\widehat{K}(u)Y(u)}\widehat{E}\{L''_{1ki}(t,u)\}^2\right], \quad (5.10)$$

where $\widehat{E}\{L_{1ki}(t,u)\}^2 = n^{-1}\sum_{i=1}^{n}\Delta_i\{Q_{ki}I(V_i \leq t) - \widehat{G}_{1k}(t,u)\}^2 I(V_i \geq u)/\widehat{K}(V_i)$, $\widehat{E}\{L'_{1ki}(t,u)\}^2 = n^{-1}\sum_{i=1}^{n}\Delta_i[Q_{ki}\{I(V_i \leq t) - \widehat{F}'_{1k}(t)\} - \widehat{G}'_{1k}(t,u)]^2 I(V_i \geq u)/\widehat{K}(V_i)$, $\widehat{G}'_{1k}(t,u) = \{n\widehat{S}(u)\}^{-1} \times \sum_{i=1}^{n}\Delta_i Q_{ki}\{I(V_i \leq t) - \widehat{F}'_{1k}(t)\}I(V_i \geq u)/\widehat{K}(V_i)$, and

$\widehat{E}\{L''_{1ki}(t,u)\}^2 = n^{-1}\sum_{i=1}^{n}\Delta_i[Q_{ki}I(V_i \le t) - \widehat{G}_{1k}(t,u) - \widehat{\alpha}_{1k}\{Q_{ki} - 1 - \widehat{G}_{Q_k}(u)\}]^2 I(V_i \geq u)/\widehat{K}(V_i)$. Variance estimators for $\widehat{\mu}_{1k}$, $\widehat{\mu}'_{1k}$, and $\widehat{\mu}''_{1k}$ are obtained by replacing $I(V_i \le t)$ by V_i and the relevant survival distribution estimators by mean estimators in the above.

Note that estimators for the survival function or mean for A_1B_1 and A_1B_2 are correlated, as both use the information from the patients who do not both remit and consent. By arguments like those for variance estimators, it is possible to derive the covariance between $\widehat{F}_{11}(t)$ and $\widehat{F}_{12}(t)$ and corresponding mean estimators, and similarly for the other estimators. These expressions are given in Appendix B and are useful for testing contrasts of survival probabilities and mean survival, as illustrated in the next section.

Chapter 6

Example and Empirical Results

6.1 Analysis of CALGB 8923

We demonstrate the proposed methods by application to the data from CALGB 8923. Because of thorough patient follow-up after the initial findings of the trial were reported and short survival times (greater than 75% mortality within two years), the current compilation of the data, which is more complete than that reported in Stone et al. (1995), does not contain appreciable censoring. Hence, to illustrate performance under the more profound censoring that would be expected at a typical time of analysis, we artificially censor the data by restricting the follow-up period to two and one-half years after the first enrollment and analyze one and one-half year restricted survival times. This reduces the sample size to 278 patients with approximately 30% censoring. For this subset of the data, approximately 50% of patients achieved remission, and of these approximately 90% were randomized to

second stage therapy. We let $j = 1, 2$ correspond to GM-CSF and placebo, and $k = 1, 2$ correspond to intensification treatments I and II, respectively.

Figure 6.1 shows survival function estimates for the four treatment policies. These do not show appreciable differences, consistent with the reported interpretation of CALGB 8923 given by Stone et al. (1995). Our methods allow explicit estimation of the survival distributions for all four therapy combinations, which may be used to assess treatment policy performance and to compare survival probabilities at specific time points. For example, consider comparing the survival distributions at one year. The three-degree-of-freedom large-sample Wald tests of $H_0 : S_{11}(365) = S_{12}(365) = S_{21}(365) = S_{22}(365)$ based on $\widehat{S}_{jk}(365)$, $\widehat{S}'_{jk}(365)$, and $\widehat{S}''_{jk}(365)$, yield approximate chi-square test statistics of 1.24 ($p = 0.74$), 1.26 ($p = 0.74$), and 0.91 ($p = 0.82$), respectively, suggesting no evidence of a difference.

Our approach also allows interpretation of the standard analysis (i). If patients are randomized to maintenance therapy with equal probability, and there are no ties, then $1 - \{\widehat{F}_{j1}(t) + \widehat{F}_{j2}(t)\}/2$ is numerically identical to the Kaplan-Meier estimator of overall survival under A_j ignoring B. Therefore, the common practice of estimating survival for the induction treatment ignoring maintenance therapy is equivalent to estimating the average of survival functions for A_jB_1 and A_jB_2 in the counterfactual framework.

Table 6.1 shows estimates of mean restricted survival time for each treatment policy. Three-degree-of-freedom tests for $H_0 : \mu_{11} = \mu_{12} = \mu_{21} = \mu_{22}$ based on $\widehat{\mu}_{jk}$, $\widehat{\mu}'_{jk}$, and $\widehat{\mu}''_{jk}$, yield test statistics and p-values of 2.67 (p=0.44), 2.73 (p=0.44), and 2.76

(p=0.43), respectively, providing no evidence of overall mean differences. All point estimates suggest GM-CSF treatment leads to a decrease in mean survival time. However, two-tailed Wald tests of $H_0 : (\mu_{11} + \mu_{12})/2 = (\mu_{21} + \mu_{22})/2$ using $\widehat{\mu}_{jk}$, $\widehat{\mu}'_{jk}$, and $\widehat{\mu}''_{jk}$, give z scores of -1.57 ($p = 0.12$), -1.56 ($p = 0.12$), and -1.53 ($p = 0.13$), respectively, indicating no strong evidence that infusions of GM-CSF influence mean survival time. Tests of $H_0 : (\mu_{11} + \mu_{21})/2 = (\mu_{12} + \mu_{22})/2$, contrasting the mean intensification treatment effect over induction therapy arms based on $\widehat{\mu}_{jk}$, $\widehat{\mu}'_{jk}$, and $\widehat{\mu}''_{jk}$ yield z scores and p-values of -0.37 (p=0.71), -0.29 (p=0.77), and -0.10 (p=0.92), indicating no evidence of a difference. Similarly, two-tailed tests for differences between intensification treatment policies within induction arms are not significant. For $H_0 : \mu_{11} = \mu_{12}$, the corresponding z values are -0.36 ($p = 0.72$), -0.33 ($p = 0.74$), and -0.03 ($p = 0.97$). For $H_0 : \mu_{21} = \mu_{22}$, the z values are -0.91 ($p = 0.36$), -0.93 ($p = 0.35$), and -0.60 ($p = 0.55$).

6.2 Simulation Results

To evaluate the performance of the proposed methods, we carried out several simulation studies. Because data from different A treatments are independent, we need simulate for only one A treatment arm. Similar to CALGB 8923, all simulations reported here were based on a two and one-half year study scenario for $n = 200$ and 500 individuals. For each individual, censoring time was generated from the uniform(0,2.5) distribution independent of all other variables. The remis-

sion/consent and B treatment indicators, R and Z, were sampled from Bernoulli(π_R) and Bernoulli(π_Z=0.5) distributions, respectively. When $R = 0$, a survival time T^*_λ was drawn from exponential(λ) with mean $1/\lambda$ and $T_{11} = T_{12}$ was taken to be the smaller of T^*_λ and $L = 1.5$, the upper bound on survival time. When $R = 1$, a remission/consent time T^*_α was drawn from exponential(α) and survival times under both B treatments were generated. Post-remission survival time under treatment B_1, T^*_{11}, was sampled from an exponential distribution with parameter e^{β_1}. Given the value of T^*_{11}, post-remission survival time under treatment B_2, T^*_{12}, was taken from an exponential distribution with parameter $e^{\beta_1+\beta_2 T^*_{11}}$. The total survival times under the B treatments for patients with $R = 1$ were then constructed as $T_{11} = \min(T^*_\alpha + T^*_{11}, 1.5)$ and $T_{22} = \min(T^*_\alpha + T^*_{12}, 1.5)$.

The simulation was conducted according to a (2×2) factorial design in choice of π_R and $E(T^*_\lambda)$. Parameter values for the exponential distributions were selected by first specifying the means of $T^*_\lambda, T^*_\alpha, T^*_{11}$, and T^*_{12} as fractions of $L = 1.5$ and solving for λ, α, β_1, and β_2. For results presented in Tables 1 and 2, the distributions of T^*_α, T^*_{11}, and T^*_{12} were held fixed. Their means as a fraction of $L = 1.5$ were 0.1, 0.5, and 1.0, respectively, giving $\alpha = 6.67$, $\beta_1 = 0.29$, and $\beta_2 = -0.67$. The two levels of $E(T^*_\lambda)/L$ were 0.3 and 0.5, corresponding to λ values of approximately 2.22 and 1.33, respectively.

For each of 1000 Monte Carlo data sets, $P(T_{1k} > t)$, $k = 1, 2$, was estimated using each of (5.1), (5.2), and (5.6) at $t = 0.5$ and $t = 1.0$, representing time points early and later in the study. The μ_{1k}, $k = 1, 2$, were estimated using the analogous expressions

for means. For each estimand, 95% Wald confidence intervals were constructed using the appropriate estimated standard errors from Section 5.1 and a critical value of 1.96.

Table 6.2 presents coverage and efficiency results for estimating the survival function, and Table 6.3 shows results for estimation of mean survival. Monte Carlo biases relative to the true values were less than 2% in almost all cases and are not displayed. From Table 6.2, interval coverage is at the nominal level for both sample sizes considered. For $n = 200$ and estimation of means in Table 6.3 , coverage is slightly lower than the nominal value; this is not surprising, as estimation of mean survival with censored data can be problematic with small sample sizes. This issue is resolved for $n = 500$, where coverage attains the nominal level. Interestingly, efficiency of (5.2) and its analog for the mean, which use probabilistically-adjusted sample size, differs. Table 6.3 shows that $\widehat{\mu}'_{1k}$ is clearly preferable to $\widehat{\mu}_{1k}$ on the basis of efficiency, whereas Table 6.2 shows $\widehat{S}_{1k}(t)$ outperforms $\widehat{S}'_{1k}(t)$ in some instances. This behavior may be explained by taking the difference in the asymptotic variances for the estimators in each case. For $\widehat{S}_{1k}(t)$ and $\widehat{S}'_{1k}(t)$, this difference may be shown to depend on the covariance of R and $I(T_{1k} \leq t) - F_{1k}(t)$, and for $\widehat{\mu}_{1k}$ and $\widehat{\mu}'_{1k}$, it depends on the covariance of R and $T_{1k} - \mu_{1k}$. Thus, a strong tendency for increased survival among patients achieving remission relative to those who do not would result in opposite signs for these covariances at some t. Our studies indicate that this phenomenon becomes more pronounced when lower remission/consent probabilities are coupled with large differences in expected survival between the remitting/consent and other popu-

lations. In any case, the simple adjustment used by $\widehat{S}''_{1k}(t)$ and $\widehat{\mu}''_{1k}$ to gain efficiency indeed offers substantial gains relative to the other two estimators, especially when estimating means, suggesting it is preferable for routine use. The slight increase in complexity is offset by improved performance.

6.3 Discussion

We have proposed methods for estimating survival distribution and mean survival time for treatment policies using data from two-stage randomization clinical trials, where the estimators have an "intent-to-treat" interpretation. Our approach addresses an important objective in the analysis of these trials for which, to our knowledge, specific methodology has heretofore been unavailable. Simulation evidence shows that the estimators offer reliable performance in realistic sample sizes. A simple adjustment to gain efficiency yields an estimator that delivers considerable improvement, and we recommend it for routine use.

Of course, an alternative approach to estimating the survival distributions or mean survival for a set of treatment policies of the form "treat with A_j followed by B_k if remission and consent" would be to randomize participants up-front to the policies and then estimate the quantities of interest using only the data observed within each group. Such an approach would yield estimates of the counterfactual survival distributions as defined here, but would be inefficient because information from nonremitting patients is not used effectively. The important issue is not the

two stages of randomization, but the way in which the estimators make use of the available data. The methods we have proposed improve efficiency of estimation of the performance of a given treatment policy by borrowing information on nonremitting patients across treatment policies.

As we have discussed, we take an "intent-to-treat" perspective, considering treatment *policies*, so that we handle data from patients who do not consent to the receive the maintenance therapy according to this principle. Our approach does not attempt to deduce causal treatment effects of regimens of the form "A_j followed by B_k if remission," taking into account the confounding role of consent to the second-stage therapy. We are currently developing methods to address this problem.

6.4 Figures and Tables

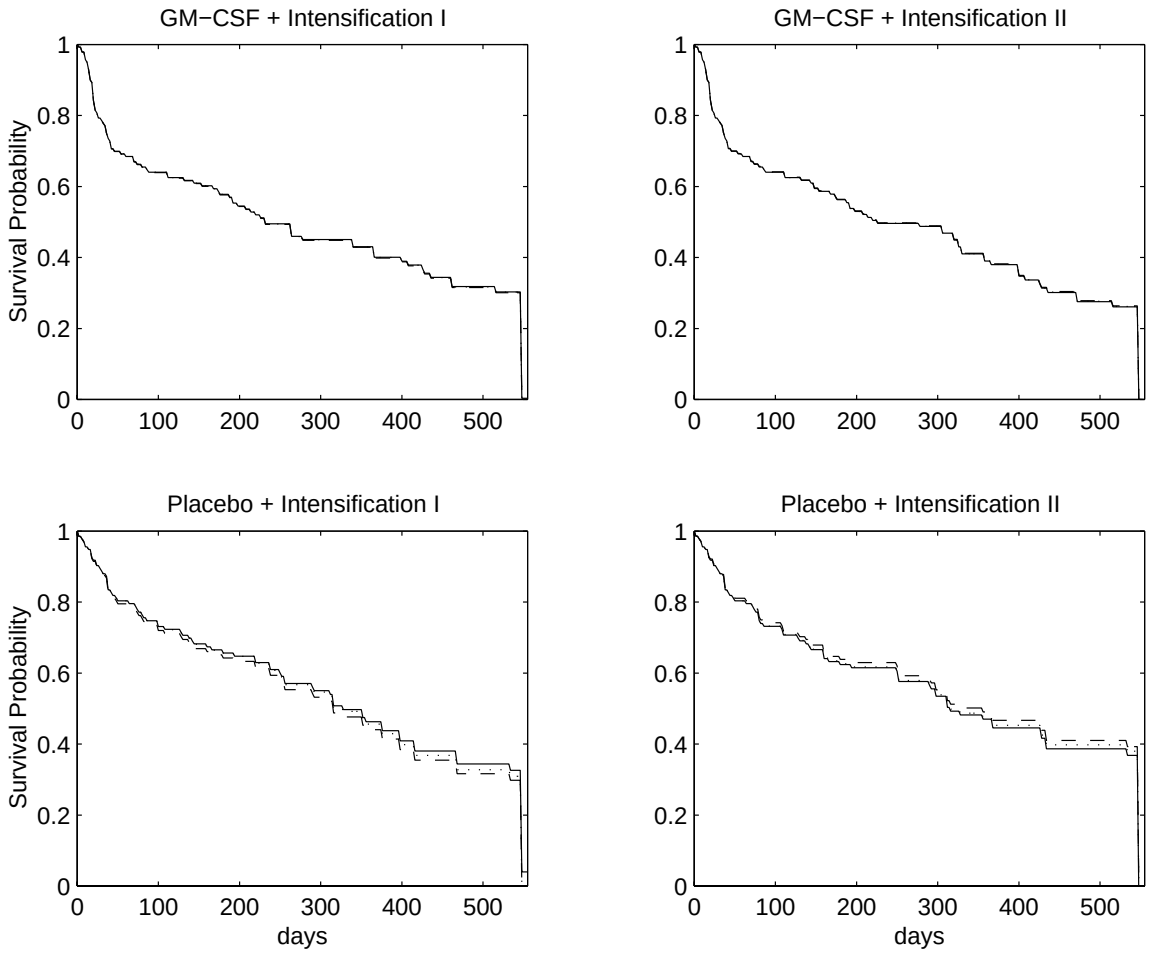

Figure 6.1: Survival function estimates for the CALGB 8923 data: $\widehat{S}(t)$ (solid line), $\widehat{S}'(t)$ (dashed line), and $\widehat{S}''(t)$ (dotted line).

Table 6.1: Estimated mean survival times for the CALGB 8923 data. Numbers in parentheses are standard error estimates.

Induction	Intensification	$\widehat{\mu}$	$\widehat{\mu}'$	$\widehat{\mu}''$
GM-CSF	I	274.7 (43.2)	275.6 (26.7)	276.4 (20.4)
	II	271.1 (42.4)	270.2 (26.4)	269.4 (20.1)
Placebo	I	292.8 (43.4)	305.0 (24.4)	309.5 (21.1)
	II	337.9 (46.6)	325.0 (26.2)	319.0 (22.3)

Table 6.2: Monte Carlo coverage and relative efficiency for estimation of survival probabilities, 1000 data sets. Entries are coverage expressed as a percentage; values in parentheses are relative efficiencies with respect to $\widehat{S}_{11}(t)$ or $\widehat{S}_{12}(t)$.

t	π_R	$E(T^*_\lambda)/L$	$S_{11}(t)$	$S_{12}(t)$	$\widehat{S}_{11}(t)$	$\widehat{S}'_{11}(t)$	$\widehat{S}''_{11}(t)$	$\widehat{S}_{12}(t)$	$\widehat{S}'_{12}(t)$	$\widehat{S}''_{12}(t)$
						$n = 200$				
0.5	0.2	0.3	0.390	0.412	94.4	94.6 (0.96)	94.4 (1.06)	95.6	94.3 (0.84)	95.4 (1.03)
		0.5	0.537	0.559	94.9	94.8 (1.01)	95.3 (1.06)	93.4	94.3 (0.97)	93.9 (1.04)
	0.5	0.3	0.481	0.534	95.0	94.7 (0.94)	94.9 (1.11)	95.4	94.1 (0.84)	94.6 (1.05)
		0.5	0.573	0.626	94.4	93.5 (1.00)	94.1 (1.09)	94.7	94.6 (1.02)	94.2 (1.09)
1.0	0.2	0.3	0.153	0.185	93.9	92.7 (1.13)	93.5 (1.33)	95.4	92.5 (0.91)	94.2 (1.14)
		0.5	0.277	0.309	94.0	94.7 (1.14)	94.4 (1.21)	93.5	93.1 (1.05)	93.6 (1.14)
	0.5	0.3	0.219	0.300	92.9	92.6 (1.47)	92.5 (1.59)	95.0	92.6 (0.94)	93.5 (1.19)
		0.5	0.296	0.378	95.9	92.5 (1.21)	93.7 (1.32)	94.9	95.7 (1.14)	95.6 (1.25)
						$n = 500$				
0.5	0.2	0.3	0.390	0.412	95.6	94.6 (0.92)	95.6 (1.06)	95.9	94.7 (0.87)	95.9 (1.03)
		0.5	0.537	0.559	95.0	94.3 (0.98)	94.4 (1.05)	94.7	95.3 (0.94)	95.4 (1.02)
	0.5	0.3	0.481	0.534	93.5	93.8 (1.02)	93.8 (1.13)	95.2	94.7 (0.91)	95.3 (1.07)
		0.5	0.573	0.626	95.9	94.9 (1.02)	95.0 (1.10)	93.6	94.1 (0.96)	93.6 (1.06)
1.0	0.2	0.3	0.153	0.185	94.0	94.6 (1.25)	94.8 (1.40)	94.5	94.2 (0.93)	93.5 (1.15)
		0.5	0.277	0.309	94.7	94.5 (1.22)	95.0 (1.27)	95.0	93.5 (1.00)	94.4 (1.10)
	0.5	0.3	0.219	0.300	94.6	95.2 (1.38)	94.8 (1.57)	94.7	94.3 (1.06)	94.9 (1.27)
		0.5	0.296	0.378	95.3	93.4 (1.39)	94.0 (1.47)	93.6	93.7 (1.15)	93.3 (1.23)

Table 6.3: Monte Carlo coverage and relative efficiency for estimation of means based on 1000 data sets. Entries are coverage expressed as a percentage; values in parentheses are relative efficiencies with respect to $\widehat{\mu}_{11}$ or $\widehat{\mu}_{12}$.

π_R	$E(T^*_\lambda)/L$	μ_{11}	μ_{12}	$\widehat{\mu}_{11}$	$\widehat{\mu}'_{11}$	$\widehat{\mu}''_{11}$	$\widehat{\mu}_{12}$	$\widehat{\mu}'_{12}$	$\widehat{\mu}''_{12}$
					$n = 200$				
0.2	0.3	0.502	0.537	92.8	92.6 (1.66)	92.6 (2.03)	93.0	92.6 (1.62)	90.8 (1.94)
	0.5	0.673	0.709	92.9	93.5 (1.72)	93.7 (1.84)	93.7	94.2 (1.81)	93.8 (1.95)
0.5	0.3	0.604	0.693	93.8	93.1 (2.17)	93.3 (2.52)	93.7	94.0 (2.31)	93.3 (2.78)
	0.5	0.711	0.800	93.1	94.0 (2.25)	93.2 (2.42)	94.8	93.7 (2.35)	93.3 (2.57)
					$n = 500$				
0.2	0.3	0.502	0.537	94.5	94.1 (1.60)	93.1 (1.87)	94.5	94.5 (1.69)	94.3 (2.22)
	0.5	0.673	0.709	94.5	95.8 (1.78)	94.9 (1.87)	93.6	93.3 (1.92)	93.5 (2.26)
0.5	0.3	0.604	0.693	95.1	95.7 (2.22)	95.2 (2.57)	94.0	94.1 (2.32)	93.1 (3.02)
	0.5	0.711	0.800	93.3	94.3 (2.32)	93.4 (2.50)	94.6	95.4 (2.78)	95.7 (3.21)

Bibliography

Carroll, R.J., Ruppert, D., and Stefanski, L.A. (1995), *Measurement Error in Nonlinear Models*, London: Chapman and Hall.

Cochran, W.G. (1968), "The Effectiveness of Adjustment by Subclassification in Removing Bias in Observational Studies," *Biometrics*, 24, 205–213.

Connors, A.F., Speroff, T., Dawson, N.V., Thomas, C., Harrell, F.E., Wagner, D., Desbiens, N., Goldman, L., Wu, A.W., Califf, R.M., Fulkerson, W.J., Vidaillet, H., Broste, S., Bellamy, P., Lynn, J., and Knaus, W.G. (1996), "The Effectiveness of Right Heart Catheterization in the Initial Care of Critically Ill Patients," *Journal of the American Medical Association*, 276, 889-897.

D'Agostino, R.B. (1998), "Tutorial in Biostatistics: Propensity Score Methods for Bias Reduction in the Comparison of a Treatment to a Non-Randomized Control Group," *Statistics in Medicine*, 17, 2265–2281.

Dawid, A.P. (1979), "Conditional independence in statistical theory," *Journal of the Royal Statistical Society, Series B.* 41, 1–31.

Dehejia, R.H. and Wahba, S, (1999), "Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs," *Journal of the American Statistical Association*, 94, 1053–1062.

Drake, C. (1993), "Effects of Misspecification of the Propensity Score on Estimators of Treatment Effects," *Biometrics*, 49, 1231–1236.

Fleming, T.R. and Harrington, D.P. (1991), *Counting Processes and Survival Analysis*, New York: John Wiley & Sons, Inc.

Hernán, Brumback, B., and Robins, J.M. (2000), "Marginal Structural Models to Estimate the Causal Effect of Zidovudine on the Survival of HIV-Positive Men," *Epidemiology*, 11, 561–570.

Holland, P.R. (1986), "Statistics and Causal Inference," *Journal of the American Statistical Association*, 81, 945–960.

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.

Huber, P.J. (1981), *Robust Statistics*, New York: John Wiley & Sons.

Joffe, M.M. and Rosenbaum, P.R. (1999), "Propensity Scores"' (invited commentary), *American Journal of Epidemiology*, 150, 327–333.

Joss, R.A., Alberto, P., Bleher, E.A., Ludwig, C., Siegenthaler, P., Martinelli, G.,

Sauter, C., Schatzmann, E., and Senn, H.J. (1994), "Combined-Modality Treatment of Small-Cell Lung Cancer: Randomized Comparison of Three Induction Chemotherapies Followed by Maintenance Chemotherapy With or Without Radiotherapy to the Chest," *Annals of Oncology*, 5, 921–928.

McIntosh, M.W. and Rubin, D.B. (1999), "On Estimating the Causal Effects of DNR Orders," *Medical Care*, 37, 722–726.

Perkins, S.M., Tu, W., Underhill, M.G., Zhou, X.-H., and Murray, M.D. (2000), "The Use of Propensity Scores in Pharmacoepidemiologic Research," *Pharmacoepidemiology and Drug Safety*, 9, 93–101.

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.

Robins, J.M., Hernán, M., and Brumback, B. (2000), "Marginal Structural Models and Causal Inference in Epidemiology," *Epidemiology*, 11, 550–560.

Rosenbaum, P.R. (1987), "Model-Based Direct Adjustment," *Journal of the American Statistical Association*, 82, 387–394.

Rosenbaum, P.R. (1998), "Propensity Score," in *Encyclopedia of Biostatistics*, New York: John Wiley and Sons.

Rosenbaum, P.R. and Rubin, D.B. (1983), "The Central Role of the Propensity Score in Observational Studies for Causal Effects," *Biometrika*, 70, 41–55.

Rosenbaum, P.R. and Rubin, D.B. (1984), " Reducing Bias in Observational Studies using Subclassification on the Propensity Score," *Journal of the American Statistical Association*, 79, 516–524.

Rubin, D.R. (1974), "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies," *Journal of Educational Psychology*, 66, 688–701.

Scharfstein, D.O., Rotnitzky, A., and Robins, J.M. (1999), Rejoinder to "Adjusting for Nonignorable Drop-Out Using Semiparametric Nonresponse Models," *Journal of the American Statistical Association*, 94, 1135–1146.

Shepardson, L.B., Youngner, S.J., Speroff, T., and Rosenthal, G.E. (1999), "Increased Risk of Death in Patients With Do-Not-Resuscitate Orders," *Medical Care*, 37, 727–737.

Stein, R.S., Vogler, W.R., Winton, E.F., Cohen, H.J., and Raney, M.R. (1990), "Therapy of Acute Meylogenous Leukemia in Patients Over the Age of 50: A Randomized Southeastern Cancer Study Group Trail," *Leukemia Research*, 14, 895–903.

Stone, R.M., Berg, D.T., George, S.L., Dodge, R.K., Paciucci, P.A., Schulman, P., Lee, E.J., Moore, J.O., Powell, B.L., and Schiffer, C.A. (1995), "Granulocyte-Macrophage Colony-Stimulating Factor After Initial Chemotherapy For Elderly Patients With Primary Acute Myelogenous Leukemia," *New England Journal of Medicine*, 332, 1671–1677.

Tummarello, D., Mari, D., Graziano, F., Isidori, P., Cetto, G., Pasini, F., Santo, A., and Cellerino, R. (1997), "A Randomized, Controlled Phase III Study of Cyclophosphamide, Doxorubicin, and Vincristine with Etoposide (CAV-E) or Teniposide (CAV-T), Followed by Recombinant Interferon-α Maintenance Therapy or Observation, in Small Cell Lung Carcinoma Patients with Complete Responses," *Cancer*, 80, 2222–2229.

Zhao, H. and Tsiatis, A.A. (1997), "A Consistent Estimator of Quality-adjusted Survival Time," *Biometrics*, 84, 339–348.

Appendix A

Applying the results in Section A.3.6 of Carroll et al. (1995) to (3.4), we have

$$\Sigma_S = A_{22}^{-1}(\boldsymbol{B}_{22} - \boldsymbol{A}_{21}\boldsymbol{A}_{11}^{-1}\boldsymbol{B}_{12} - \boldsymbol{B}_{12}^T\boldsymbol{A}_{11}^{-T}\boldsymbol{A}_{21}^T + \boldsymbol{A}_{21}\boldsymbol{A}_{11}^{-1}\boldsymbol{B}_{11}\boldsymbol{A}_{11}^{-T}\boldsymbol{A}_{21}^T)A_{22}^{-T}, \quad \text{(A.1)}$$

where the matrices in this expression follow from tedious evaluation of the required derivatives and covariance matrix. In particular, it may be shown that $A_{22} = -1$, and

$$\boldsymbol{A}_{11} = \begin{pmatrix} \boldsymbol{E}_{qq} & \boldsymbol{0} & \boldsymbol{E}_{q\beta} \\ \boldsymbol{E}_{pq} & -\boldsymbol{I}_K & \boldsymbol{E}_{p\beta} \\ \boldsymbol{0} & \boldsymbol{0} & -\boldsymbol{E}_{\beta\beta} \end{pmatrix}, \quad \boldsymbol{B}_{11} = \begin{pmatrix} \boldsymbol{F}_{qq} & \boldsymbol{F}_{qp} & \boldsymbol{0} \\ \boldsymbol{F}_{qp}^T & \boldsymbol{F}_{pp} & \boldsymbol{F}_{p\beta} \\ \boldsymbol{0} & \boldsymbol{F}_{p\beta}^T & \boldsymbol{E}_{\beta\beta} \end{pmatrix}.$$

Here, $\boldsymbol{E}_{qq} = \text{diag}\{f_e(q_1), f_e(q_2), \ldots, f_e(q_{K-1})\}$; $\boldsymbol{E}_{pq}^{(i,j)} = q_j f_e(q_j)$, $i = j$, $-q_j f_e(q_j)$, $i = j+1$, and zero otherwise $(K \times K-1)$; and $\boldsymbol{E}_{q\beta}$ $(K-1 \times p)$ has jth row $\partial/\partial\boldsymbol{\beta}^T\{\int_0^{q_j} f_e(t)dt\}$ and $\boldsymbol{E}_{p\beta}$ $(K \times p)$ has jth row $\partial/\partial\boldsymbol{\beta}^T\{\int_{q_{j-1}}^{q_j} tf_e(t)dt\}$, where differentiation is with respect to $\boldsymbol{\beta}$ in $f_e(\cdot)$ only. In addition, $\boldsymbol{F}_{qq}$ is symmetric with (i,j) upper-triangular element $(i/K)(1-j/K)$; $\boldsymbol{F}_{qp}^{(i,j)} = p_j(1-i/K)$, $i \geq j$, $= -p_j(i/K)$, $i < j$ $(K-1 \times K)$; $\boldsymbol{F}_{pp}$ $(K \times K)$ is symmetric with $\boldsymbol{F}_{pp}^{(j,j)} = p_j(1-p_j)$, $\boldsymbol{F}_{pp}^{(i,j)} = -p_i p_j$; and $\boldsymbol{F}_{p\beta}$ $(K \times p)$ has jth row $E\{I(e \in Q_j)e_\beta^T\}$, where the expectation is with respect to the distribution of $\boldsymbol{X}$. Defining $h_{1j} = p_j^{-1}\int_{q_{j-1}}^{q_j} E(Y_1|t)tf_e(t)dt$ and $h_{0j} =$

$(1/K-p_j)^{-1}\int_{q_{j-1}}^{q_j} E(Y_0|t)(1-t)f_e(t)dt$, $j=1,\ldots,K$, and $g_{1j}=E(Y_1|q_j)q_j(p_j^{-1}-p_{j+1}^{-1})$ and $g_{0j}=E(Y_0|q_j)(1-q_j)\{(1/K-p_j)^{-1}-(1/K-p_{j+1})^{-1}\}$, $j=1,\ldots,K-1$, then $\boldsymbol{A}_{21}=(\ \boldsymbol{E}_{\Delta q}\ \ \boldsymbol{E}_{\Delta p}\ \ \boldsymbol{E}_{\Delta\beta}\ )$, $\boldsymbol{B}_{12}^T=(\ \boldsymbol{F}_{q\Delta}^T\ \ \boldsymbol{F}_{p\Delta}^T\ \ \boldsymbol{F}_{\beta\Delta}^T\ )^T$, where $\boldsymbol{E}_{\Delta p}$ $(1\times K)$ has jth element $(p_jK)^{-1}h_{1j}-(1-Kp_j)^{-1}h_{0j}$, respectively; $\boldsymbol{E}_{\Delta q}$ $(1\times K-1)$ has elements $K^{-1}(g_{1j}-g_{0j})f_e(q_j)$; and $\boldsymbol{E}_{\Delta\beta}$ $(1\times p)$ is given by

$$\partial/\partial\boldsymbol{\beta}^T[\sum_{j=1}^{K}\{(p_jK)^{-1}\int_{q_{j-1}}^{q_j}E(Y_1|t)tf_e(t)dt-(K^{-1}-p_j)^{-1}\int_{q_{j-1}}^{q_j}E(Y_0|t)(1-t)f_e(t)dt\}],$$

where differentiation is with respect to $\boldsymbol{\beta}$ in $f_e(\cdot)$. Similarly, $\boldsymbol{F}_{p\Delta}^T$ $(1\times K)$ has jth element $K^{-1}h_{1j}-p_j\Delta^*$; $\boldsymbol{F}_{q\Delta}^T$ $(1\times K-1)$ has elements $K^{-1}\sum_{i=1}^{j}(h_{1i}-h_{0i}-\Delta^*)$; and $\boldsymbol{F}_{\beta\Delta}^T$ $(1\times p)$ is $K^{-1}\sum_{j=1}^{K}[p_j^{-1}E\{Y_1I(e\in Q_j)e_\beta^T\}+(1/K-p_j)^{-1}E\{Y_0I(e\in Q_j)e_\beta^T\}$.

Substituting these expressions in (A.1) and simplifying yields (3.5), with $\Gamma_p=\boldsymbol{E}_{\Delta p}\boldsymbol{F}_{p\Delta}+\boldsymbol{F}_{p\Delta}^T\boldsymbol{E}_{\Delta p}^T+\boldsymbol{E}_{\Delta p}F_{pp}\boldsymbol{E}_{\Delta p}^T$, $\Gamma_{qp}=-\boldsymbol{H}_{\Delta q}(\boldsymbol{E}_{\Delta p}\boldsymbol{F}_{qp}^T+\boldsymbol{F}_{q\Delta}^T)^T-(\boldsymbol{E}_{\Delta p}\boldsymbol{F}_{qp}^T+\boldsymbol{F}_{q\Delta}^T)\boldsymbol{H}_{\Delta q}^T+\boldsymbol{H}_{\Delta q}\boldsymbol{F}_{qq}\boldsymbol{H}_{\Delta q}^T$, and $\Gamma_{\beta qp}=(\boldsymbol{H}_{\Delta\beta}-\boldsymbol{H}_{\Delta q}\boldsymbol{E}_{q\beta})\boldsymbol{E}_{\beta\beta}^{-1}(F_{\beta\Delta}^T+\boldsymbol{E}_{\Delta p}\boldsymbol{F}_{\beta p}^T)^T+(F_{\beta\Delta}^T+\boldsymbol{E}_{\Delta p}\boldsymbol{F}_{\beta p}^T)\boldsymbol{E}_{\beta\beta}^{-1}(\boldsymbol{H}_{\Delta\beta}-\boldsymbol{H}_{\Delta q}\boldsymbol{E}_{q\beta})^T+(\boldsymbol{H}_{\Delta\beta}-\boldsymbol{H}_{\Delta q}\boldsymbol{E}_{q\beta})\boldsymbol{E}_{\beta\beta}^{-1}(\boldsymbol{H}_{\Delta\beta}-\boldsymbol{H}_{\Delta q}\boldsymbol{E}_{q\beta})^T$, where $\boldsymbol{H}_{\Delta q}=(\boldsymbol{E}_{\Delta q}+\boldsymbol{E}_{\Delta p}\boldsymbol{E}_{pq})\boldsymbol{E}_{qq}^{-1}$ and $\boldsymbol{H}_{\Delta\beta}=\boldsymbol{E}_{\Delta\beta}+\boldsymbol{E}_{\Delta p}\boldsymbol{E}_{p\beta}$.

To obtain (3.8), note that the estimating equation for Δ corresponding to (2.7) is $\sum_{i=1}^{n}\psi_\Delta^{IPW1}(Y_i,Z_i,\boldsymbol{X}_i,\boldsymbol{V}_i,\Delta,\boldsymbol{\beta},\boldsymbol{\gamma})=0$, where $\psi_\Delta^{IPW1}(Y,Z,\boldsymbol{X},\boldsymbol{V},\Delta,\boldsymbol{\beta},\boldsymbol{\gamma})=ZY/e(\boldsymbol{X},\boldsymbol{V},\boldsymbol{\beta},\boldsymbol{\gamma})-(1-Z)Y/\{1-e(\boldsymbol{X},\boldsymbol{V},\boldsymbol{\beta},\boldsymbol{\gamma})\}-\Delta$. We may thus write the full set of equations as

$$\sum_{i=1}^{n}\{\Psi_{\gamma\beta}^T(Z_i,\boldsymbol{X}_i,\boldsymbol{\beta},\boldsymbol{\gamma}),\psi_\Delta^{IPW1}(Y_i,Z_i,\boldsymbol{X}_i,\boldsymbol{V}_i,\Delta,\boldsymbol{\beta},\boldsymbol{\gamma})\}=\boldsymbol{0},$$

where $\Psi_{\gamma\beta}$ represents (3.7) and the modified version of (2.4) stacked. From Section A.3.6 of Carroll et al. (1995), Σ_{IPW1}^V may be written in the same form as (A.1), where

again $A_{22} = -1$. Equation (3.8) follows from evaluation of the remaining elements of this expression and simplification.

To obtain the second term in (3.12), let $\boldsymbol{E}_{\Delta\gamma}$ $(1 \times q)$ equal

$$\partial/\partial\boldsymbol{\gamma}^T[\sum_{j=1}^{K}\{(p_jK)^{-1}\int_{q_{j-1}}^{q_j} E(Y_1|t)tf_e(t)dt - (K^{-1}-p_j)^{-1}\int_{q_{j-1}}^{q_j} E(Y_0|t)(1-t)f_e(t)dt\}].$$

Let $\boldsymbol{E}_{q\gamma}$ $(K-1 \times q)$ and $\boldsymbol{E}_{p\gamma}$ $(K \times q)$ have jth rows $\partial/\partial\boldsymbol{\gamma}^T\{\int_0^{q_j} f_e(t)dt\}$ and $\partial/\partial\boldsymbol{\gamma}^T\{\int_{q_{j-1}}^{q_j} tf_e(t)dt\}$, respectively. Also let $\boldsymbol{F}_{p\gamma}$ $(K \times q)$ be the matrix with jth row $E\{I(e \in Q_j)e_\gamma^T\}$, and $\boldsymbol{F}_{\gamma\Delta}^T$ $(1 \times p)$ is $K^{-1}\sum_{j=1}^{K}[p_j^{-1}E\{Y_1I(e \in Q_j)e_\gamma^T\} + (1/K - p_j)^{-1}E\{Y_0I(e \in Q_j)e_\gamma^T\}$. Defining $\boldsymbol{H}_{\Delta\gamma} = \boldsymbol{E}_{\Delta\gamma} - \boldsymbol{E}_{\Delta p}\boldsymbol{E}_{p\gamma}$, $\boldsymbol{D}_\gamma = \boldsymbol{H}_{\Delta\gamma} - \boldsymbol{H}_{\Delta\beta}\boldsymbol{E}_{\beta\beta}^{-1}\boldsymbol{E}_{\gamma\beta}^T - \boldsymbol{H}_{\Delta q}(\boldsymbol{E}_{q\gamma} - \boldsymbol{E}_{q\beta}\boldsymbol{E}_{\beta\beta}^{-1}\boldsymbol{E}_{\gamma\beta}^T)$, and $\boldsymbol{G}_\gamma = (\boldsymbol{F}_{\gamma\Delta} - \boldsymbol{E}_{\gamma\beta}\boldsymbol{E}_{\beta\beta}^{-1}\boldsymbol{F}_{\beta\Delta})^T + \boldsymbol{E}_{\Delta p}(\boldsymbol{F}_{p\gamma}^T - \boldsymbol{E}_{\gamma\beta}\boldsymbol{E}_{\beta\beta}^{-1}\boldsymbol{E}_{\beta p})^T$, one can show that $\Gamma_{\gamma\beta qp} = \boldsymbol{D}_\gamma\boldsymbol{H}_{\gamma\beta}^{-1}\boldsymbol{G}_\gamma^T + \boldsymbol{G}_\gamma\boldsymbol{H}_{\gamma\beta}^{-1}\boldsymbol{D}_\gamma^T + \boldsymbol{D}_\gamma\boldsymbol{H}_{\gamma\beta}^{-1}\boldsymbol{D}_\gamma^T$.

Appendix B

We sketch arguments to show consistency and asymptotic normality for the proposed estimators of $F_{1k}(t)$; arguments for the estimators of the counterfactual means are analogous. Consistency of $\widehat{F}_{1k}(t)$ and $\widehat{F}'_{1k}(t)$ may be deduced straightforwardly by writing

$$\begin{aligned}\widehat{F}_{1k}(t) - F_{1k}(t) &= n^{-1}\sum_{i=1}^{n}\left\{\frac{\Delta_i Q_{ki}}{K(T_i)}I(T_i \le t) - F_{1k}(t)\right\} \\ &\quad -n^{-1}\sum_{i=1}^{n}\{\Delta_i Q_{ki} I(T_i \le t)\}\left\{\frac{\widehat{K}(T_i) - K(T_i)}{\widehat{K}(T_i)K(T_i)}\right\}, \quad \text{(B.2)}\end{aligned}$$

$$\widehat{F}'_{1k}(t) - F_{1k}(t) = \left\{n^{-1}\sum_{i=1}^{n}\frac{\Delta_i Q_{ki}}{\widehat{K}(V_i)}\right\}^{-1} n^{-1}\sum_{i=1}^{n}\frac{\Delta_i Q_{ki}}{\widehat{K}(V_i)}\{I(T_i \le t) - F_{1k}(t)\}. \quad \text{(B.3)}$$

That the right hand side of each of (B.2) and (B.3) is $o_p(1)$ may be established using arguments similar to those in the Appendix of Zhao and Tsiatis (1997).

To derive the large sample distributions of $n^{1/2}\{\widehat{F}_{1k}(t) - F_{1k}(t)\}$ and $n^{1/2}\{\widehat{F}'_{1k}(t) - F_{1k}(t)\}$, define the filtration $\mathcal{F}_n(t)$ as the increasing sequence of sub σ-algebras $\sigma\{I(C_i \le u), u \le t; R_i, Z_i, T_{11i}, T_{12i},\ i = 1 \ldots n\}$, containing all observed censoring and survival information up to time t and all information on the counterfactual survival times. Letting $\lambda^c(u)$ be the hazard function for the censoring distribution,

the corresponding martingale process is $M_i^c(t) = N_i^c(t) - \int_0^t \lambda^c(u) Y_i(u) du$, where $N_i^c(t) = I(V_i \le t, \Delta_i = 0)$ and $Y_i(u) = I(V_i \ge u)$. With $M^c(u) = \sum_{i=1}^n M_i^c(u)$ and $Y(u) = \sum_{i=1}^n Y_i(u)$, (A.4)–(A.6) of Zhao and Tsiatis (1997) yield

$$\begin{aligned} & n^{1/2}\{\widehat{F}_{1k}(t) - F_{1k}(t)\} \\ & \quad = n^{-1/2} \sum_{i=1}^n \{Q_{ki} I(T_i \le t) - F_{1k}(t)\} \\ & \qquad -n^{-1/2} \sum_{i=1}^n \int_0^L \frac{dM_i^c(u)}{K(u)} \{Q_{ki} I(T_i \le t) - \widehat{G}_{1k}(t,u)\} \\ & \quad = n^{-1/2} \sum_{i=1}^n \{Q_{ki} I(T_i \le t) - F_{1k}(t)\} \\ & \qquad -n^{-1/2} \sum_{i=1}^n \int_0^L \frac{dM_i^c(u)}{K(u)} \{Q_{ki} I(T_i \le t) - G_{1k}(t,u)\} + o_p(1) \end{aligned} \tag{B.4}$$

where $\widehat{G}_{1k}(t,u) = \{n\widehat{S}(u)\}^{-1} \sum_{i=1}^n \Delta_i Q_{ki} I(T_i \le t) I(T_i \ge u)/\widehat{K}(T_i)$ and $G_{1k}(t,u) = E\{I(T_{1ki} \le t) I(T_{1ki} \ge u)\}/P(T_i > u) = P(u \le T_{1k} \le t)/P(T > u)$; (B.4) follows by arguments similar to those in Zhao and Tsiatis (1997). Thus, from (B.4), $n^{1/2}\{\widehat{F}_{1k}(t) - F_{1k}(t)\} = n^{-1/2} \sum_{i=1}^n \psi_{ki} + o_p(1)$, where

$$\psi_{ki} = Q_{ki} I(T_i \le t) - F_{1k}(t) - \int_0^L \frac{Q_{ki} I(T_i \le t) - G_{1k}(t,u)}{K(u)} dM_i^c(u). \tag{B.5}$$

Thus, $\widehat{F}_{1k}(t)$ is an asymptotically linear estimator; i.e. $n^{1/2}$ times the estimator minus $F_{1k}(t)$ is equal to $n^{-1/2}$ times the sum of i.i.d. mean-zero random variables plus a term of $o_p(1)$, where a term in the sum is referred to as the influence function of the estimator. It follows that the influence function for $\widehat{F}_{1k}(t)$ is given by (B.5). The asymptotic variance of the estimator is the variance of the influence function, a fact we make use of momentarily.

Because $Q_{ki}I(T_i \leq t) - F_{1k}(t)$ has mean zero and is $\mathcal{F}(0)$ measurable, and the second component of (B.5) has mean zero and is uncorrelated with the first, it follows that $n^{1/2}\{\widehat{F}_{1k}(t) - F_{1k}(t)\}$ converges to a mean-zero normal distribution with variance

$$E(\psi_{ki})^2 \;=\; E\left\{Q_{ki}I(T_i \leq t) - F_{1k}(t)\right\}^2 + \int_0^L \frac{E\{L_{1ki}(t,u)\}^2}{K(u)}\lambda^c(u)du,$$

where $L_{1ki}(t,u) = \{Q_{ki}I(T_i \leq t) - G_{1k}(t,u)\}I(T_i \geq u)$. This variance may be estimated by (5.8). An entirely similar argument may be used to show that $n^{1/2}\{\widehat{F}'_{1k}(t) - F_{1k}(t)\} = n^{-1/2}\sum_{i=1}^n \psi'_{ki} + o_p(1)$, where the influence function is given by

$$\psi'_{ki} = Q_{ki}\{I(T_i \leq t) - F_{1k}(t)\} - \int_0^L \frac{Q_{ki}\{I(T_i \leq t) - F_{1k}(t)\} - G'_{1k}(t,u)}{K(u)} dM_i^c(u), \tag{B.6}$$

with $G'_{1k}(t,u) = E[\{I(T_{1ki} \leq t) - F_{1k}(t)\}I(T_{1ki} \geq u)]/P(T_i > u)$, leading to the variance

$$E(\psi'_{ki})^2 \;=\; E\left[Q_{ki}\{I(T_i \leq t) - F_{1k}(t)\}\right]^2 + \int_0^L \frac{E\{L'_{1ki}(t,u)\}^2}{K(u)}\lambda^c(u)du,$$

where $L'_{1ki}(t,u) = \left[Q_{ki}\{I(T_i \leq t) - F_{1k}(t)\} - G'_{1k}(t,u)\right] I(T_i \geq u)$. This variance may be estimated by (5.9).

To derive (5.6), the influence function corresponding to (5.5) for fixed α_{1k}, using (B.5) and (B.6), is given by

$$\psi_{ki} - \alpha_{1k}\left\{Q_{ki} - 1 - \int_0^L \frac{Q_{ki} - 1 - G_{Q_k}(u)}{K(u)} dM_i^c(u)\right\}, \tag{B.7}$$

with $G_{Q_k}(u) = E\{(Q_{ki} - 1)I(T_i \geq u)\}/P(T_i > u)$; thus, α_{1k} should be chosen to minimize the variance

$$E\left[\psi_{ki} - \alpha_{1k}\left\{Q_{ki} - 1 - \int_0^L \frac{Q_{ki} - 1 - G_{Q_k}(u)}{K(u)} dM_i^c(u)\right\}\right]^2,$$

which yields

$$\alpha_{1k} = \frac{E\{Q_{ki}(Q_{ki}-1)I(T_i \le t)\} + \int_0^L \lambda^c(u)K(u)^{-1}E\{L_{1k}^{\alpha}(t,u)\}du}{E(Q_{ki}-1)^2 + \int_0^L \lambda^c(u)K(u)^{-1}E\{G_k^{\alpha}(u)\}du}, \tag{B.8}$$

where $E\{L_{1k}^{\alpha}(u)\} = E[\{Q_{ki}I(T_i \le t) - G_{1k}(t,u)\}\{Q_{ki} - 1 - G_{Q_k}(u)\}I(T_i \ge u)]$, and $E\{G_k^{\alpha}(t,u)\} = E[\{Q_{ki} - 1 - G_{Q_k}(u)\}^2 I(T_i \ge u)]$, leading to (5.6).

Consistency of $\widehat{F}''_{1k}(t)$ follows by noting

$$\begin{aligned}
\widehat{F}''_{1k}(t) - F_{1k}(t) &= \widehat{F}_{1k}(t) - F_{1k}(t) - \widehat{\alpha}_{1k}\left\{n^{-1}\sum_{i=1}^{n}\frac{\Delta_i}{\widehat{K}(T_i)}(Q_{ki}-1)\right\} \\
&= \widehat{\alpha}_{1k}\left\{n^{-1}\sum_{i=1}^{n}\frac{\Delta_i}{\widehat{K}(T_i)}(Q_{ki}-1)\right\} + o_p(1).
\end{aligned} \tag{B.9}$$

The expectations in (B.8) can be consistently estimated as in (5.7), and the term in braces may be shown to converge in probability to zero by arguments similar to those used to show the right hand sides of (B.2) and (B.3) are $o_p(1)$. Moreover,

$$\begin{aligned}
n^{1/2}\left\{\widehat{F}''_{1k}(t) - F_{1k}(t)\right\} &= n^{1/2}\left\{\widehat{F}_{1k}(t) - F_{1k}(t)\right\} - \alpha_{1k}\left\{n^{-1/2}\sum_{i=1}^{n}\frac{\Delta_i}{\widehat{K}(T_i)}(Q_{ki}-1)\right\} \\
&\quad + (\alpha_{1k} - \widehat{\alpha}_{1k})\left\{n^{-1/2}\sum_{i=1}^{n}\frac{\Delta_i}{\widehat{K}(T_i)}(Q_{ki}-1)\right\}.
\end{aligned}$$

The same techniques used to obtain (B.4) can be used to show the term in braces is an asymptotically linear estimator of zero with influence function equal to the braced term in (B.7). Because $(\alpha_{1k} - \widehat{\alpha}_{1k})$ converges in probability to zero, estimating α_{1k} has no effect asymptotically. Hence, ψ''_{ki}, the influence function of $\widehat{F}''_{1k}(t)$, is given by (B.7) substituting (B.8), and has variance

$$E(\psi''_{ki})^2 = E\left\{Q_{ki}I(T_i \le t) - F_{1k}(t) - \alpha_{1k}(Q_{ki}-1)\right\}^2 + \int_0^L \frac{E\{L''_{1ki}(t,u)\}^2}{K(u)}\lambda^c(u)du,$$

where $L''_{1ki}(t,u) = [Q_{ki}I(T_i \le t) - G_{1k}(t,u) - \alpha_{1k}\{Q_{ki} - 1 - G_{Q_k}(u)\}]\, I(T_i \ge u)$. This variance may be estimated by (5.10).

Large sample covariances for the proposed estimators for $S_{11}(t)$ and $S_{12}(t)$ are given by the expectation of the product of the corresponding influence functions, and are given by

$$\begin{aligned} E(\psi_{1i}\psi_{2i}) &= E\{Q_{1i}Q_{2i}I(T_i \le t)\} - F_{11}(t)F_{12}(t) \\ &\quad + \int_0^L [E\{L_{11i}(t,u)L_{12i}(t,u)\}/K(u)]\lambda^c(u)du, \end{aligned}$$

$$\begin{aligned} E(\psi'_{1i}\psi'_{2i}) &= E\left[Q_{1i}Q_{2i}\{I(T_i \le t) - F_{11}(t)\}\{I(T_i \le t) - F_{12}(t)\}\right] \\ &\quad + \int_0^L [E\{L'_{11i}(t,u)L'_{12i}(t,u)\}/K(u)]\lambda^c(u)du, \end{aligned}$$

and

$$\begin{aligned} &E(\psi''_{1i}\psi''_{2i}) \\ &= E\left[\{Q_{1i}I(T_i \le t) - F_{11}(t) - \alpha_{11}(Q_{1i} - 1)\}\{Q_{2i}I(T_i \le t) - F_{12}(t) - \alpha_{12}(Q_{2i} - 1)\}\right] \\ &\quad + \int_0^L [E\{L''_{11i}(t,u)L''_{12i}(t,u)\}/K(u)]\lambda^c(u)du. \end{aligned}$$

Estimators for these covariances are

$$\begin{aligned} \widehat{\mathrm{cov}}\{\widehat{S}_{11}(t), \widehat{S}_{12}(t)\} &= n^{-1}\left[n^{-1}\sum_{i=1}^n \frac{\Delta_i Q_{1i}Q_{2i}}{\widehat{K}(V_i)} I(V_i \le t) - \widehat{F}_{11}(t)\widehat{F}_{12}(t) \right. \\ &\quad \left. + \int_0^L \frac{dN^c(u)}{\widehat{K}(u)Y(u)} \widehat{E}\{L_{11i}(t,u)L_{12i}(t,u)\}\right], \end{aligned}$$

$$\begin{aligned} \widehat{\mathrm{cov}}\{\widehat{S}'_{11}(t), \widehat{S}'_{12}(t)\} &= n^{-1}\left[n^{-1}\sum_{i=1}^n \frac{\Delta_i Q_{1i}Q_{2i}}{\widehat{K}(V_i)} \{I(V_i \le t) - \widehat{F}'_{11}(t)\}\{I(V_i \le t) - \widehat{F}'_{12}(t)\} \right. \\ &\quad \left. + \int_0^L \frac{dN^c(u)}{\widehat{K}(u)Y(u)} \widehat{E}\{L'_{11i}(t,u)L'_{12i}(t,u)\}\right], \end{aligned}$$

and

$$\widehat{\mathrm{cov}}\{\widehat{S}''_{11}(t),\widehat{S}''_{12}(t)\} = n^{-1}\left[n^{-1}\sum_{i=1}^{n}\frac{\Delta_i}{\widehat{K}(V_i)}\{Q_{1i}I(V_i\le t)-\widehat{F}''_{11}(t)-\widehat{\alpha}_{11}(Q_{1i}-1)\}\times \right.$$
$$\left.\{Q_{2i}I(V_i\le t)-\widehat{F}''_{12}(t)-\widehat{\alpha}_{12}(Q_{2i}-1)\}+\int_0^L\frac{dN^c(u)}{\widehat{K}(u)Y(u)}\widehat{E}\{L''_{11i}(t,u)L''_{12i}(t,u)\}\right],$$

where

$$\widehat{E}\{L_{11i}(t,u)L_{12i}(t,u)\}$$
$$= n^{-1}\sum_{i=1}^{n}\Delta_i\{Q_{1i}I(V_i\le t)-\widehat{G}_{11}(t,u)\}\{Q_{2i}I(V_i\le t)-\widehat{G}_{12}(t,u)\}I(V_i\ge u)/\widehat{K}(V_i),$$

$$\widehat{E}\{L'_{11i}(t,u)L'_{12i}(t,u)\} = n^{-1}\sum_{i=1}^{n}\Delta_i[Q_{1i}\{I(V_i\le t)-\widehat{F}'_{11}(t)\}-\widehat{G}'_{11}(t,u)]$$
$$\times[Q_{2i}\{I(V_i\le t)-\widehat{F}'_{12}(t)\}-\widehat{G}'_{12}(t,u)]I(V_i\ge u)/\widehat{K}(V_i),$$

and

$$\widehat{E}\{L''_{11i}(t,u)L''_{12i}(t,u)\}$$
$$= n^{-1}\sum_{i=1}^{n}\Delta_i[Q_{1i}I(V_i\le t)-\widehat{G}''_{11}(t,u)-\widehat{\alpha}_{11}\{Q_{1i}-1-\widehat{G}_{Q_1}(u)\}]$$
$$\times[Q_{2i}I(V_i\le t)-\widehat{G}''_{12}(t,u)-\widehat{\alpha}_{12}\{Q_{2i}-1-\widehat{G}_{Q_2}(u)\}]I(V_i\ge u)/\widehat{K}(V_i).$$

Substitution of V_i for $I(V_i\le t)$ and replacing estimates of $F_{1k}(t)$ with the appropriate estimates of μ_{1k} yields covariance estimators for mean survival.