Abstract

LEON, SELENE REYES. Semiparametric Efficient Estimation of Treatment Effect in a Pretest-Posttest Study with Missing Data. (Under the direction of Anastasios A. Tsiatis and Marie Davidian.)

Inference on treatment effect in a pretest-posttest study is a routine objective in medicine, public health, and other fields, and a number of approaches have been advocated. Typically, subjects are randomized to two treatments, the response is measured at baseline and a prespecified follow-up time, and interest focuses on the effect of treatment on follow-up mean response. Covariate information at baseline and in the intervening period until follow-up may also be collected. Missing posttest response for some subjects is routine, and disregarding these missing cases can lead to biased and inefficient inference. Despite the widespread popularity of this design, a consensus on an appropriate method of analysis when no data are missing, let alone on an accepted practice for taking account of missing follow-up response, does not exist.

We take a semiparametric perspective, making no assumptions about the distributions of baseline and posttest responses. Exploiting the work of Robins et al. (1994), we characterize the class of all consistent estimators for treatment effect, identify

the efficient member of this class, and propose practical procedures for implementation. The result is a unified framework for handling pretest-posttest inferences when follow-up response may be missing at random that allows the analyst to incorporate baseline and intervening information so as to improve efficiency of inference. Simulation studies and application to data from an HIV clinical trial illustrate the utility of the approach.

SEMIPARAMETRIC EFFICIENT ESTIMATION OF TREATMENT EFFECT IN A PRETEST-POSTTEST STUDY WITH MISSING DATA

by

SELENE LEON

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

2003

APPROVED BY:

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

To my parents and my sisters.

Biography

Selene Leon was born and grew up in Mexico City, Mexico. She obtained her undergraduate degree in Applied Mathematics at the Technological Institute of Mexico (ITAM) and came to North Carolina to join Graduate School at North Carolina State University. In 2001, she earned a Master degree in Statistics with Mathematical Statistic Concentration and currently is a candidate for a doctoral degree in Statistics.

Acknowledgements

I am grateful to so many persons and organizations that enumerating them all would be another chapter of this dissertation. So, missing someone is just due to lack of space, not of memory.

The support and patience that my advisors, Dr. Tsiatis and Dr. Davidian, consistently showed during my doctoral research is an inspiration to continue their example, under a different context, in my professional development.

I am also grateful to all faculty members from the Statistics Department and other Departments for their willingness to provide skills that open opportunities to persons who consider that scientific knowledge is nowadays, a fundamental tool for professional development. I also truly appreciate the friendship, encouragement and moral support from all faculty, staff and students from NC State U.

I want to thank Michael Hughes, Heather Gorski, and the AIDS Clinical Trials Group for providing the data from ACTG 175.

I am also grateful for the economical support provided by the Graduate School, as well as the grants R01-CA51962, R01-CA085848, R01-AI31789 from the National Institutes of Health for my doctoral studies. I also want to thank GlaxoSmithKline Inc. for the opportunity to practice the skills and scientific knowledge acquired in school in an industrial context.

I also want to thank my parents, specially my mother, my sisters Lluvia and Elizabeth, for their endless support and love. I also thank Emmanuel for his encouragement and love.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The pretest-posttest trial is ubiquitous in research in medicine, public health, and numerous other felds. In the usual study, subjects are randomized to one of two treatments (e.g., treatment and control) and the response of interest is ascertained for each at baseline (pre-treatment) and follow-up. The objective is to evaluate whether treatment affects follow-up response, with baseline responses serving as a basis for comparison. For instance, in many HIV clinical trials, interest focuses on comparing treatment effects on viral load or CD4 count after a specified period, with baseline observations on these quantities routinely available.

A number of strategies have been advocated for evaluation of treatment effect in this setting, including the two-sample t-test comparing follow-up observations in each group, ignoring baseline; the paired t-test comparing differences between follow-up and baseline responses; and analysis of covariance procedures applied to the follow-up responses, with either baseline response only or baseline and its interaction with

treatment included as a covariates in a linear model, referred to by Yang and Tsiatis (2001) as ANCOVA I and ANCOVA II, respectively. A nonparametric approach in a different spirit was proposed by Quade (1982); we do not consider this here. The two-sample t-test seems predicated on the assumption that baseline and follow-up responses are uncorrelated, so that no precision is to be gained from incorporating baseline information, while the others (paired t-test and ANCOVA I and II) implicitly rely on an apparent assumption of linear dependence between follow-up and baseline response. All are often associated with the assumption of normality.

Several authors (e.g., Brogan and Kutner, 1980; Crager, 1987; Laird, 1983; Stanek, 1988; Stein, 1989; Follmann, 1991) have studied these approaches under various assumptions, including normality or equality of variances of baseline and follow-up responses. Despite this work and the widespread interest in this problem, there is still no consensus on which approach is preferable under general conditions; in our experience with HIV clinical trials the paired t-test is often used because "interest focuses on differences," with no theoretical justification. An attempt to address this issue was presented by Yang and Tsiatis (2001), who studied the large-sample properties of treatment effect estimators based on these approaches under very general conditions where only the first and second moments of baseline and follow-up responses exist and may differ and their joint distribution conditional on treatment may be arbitrary. They also considered a "generalized estimating equation" (GEE) approach (e.g., Singer and Andrade, 1997) where baseline and follow-up data are treated as a multivariate response with arbitrary mean and covariance matrix. Yang and Tsiatis

(2001) showed that all these estimators are consistent and asymptotically normal; the GEE estimator is asymptotically equivalent to that from ANCOVA II and most efficient; when the randomization probability is 0.5 or covariance between baseline and follow-up responses is the same for both treatment and control, the ANCOVA I estimator is asymptotically equivalent to ANCOVA II/GEE; and, if baseline and follow-up responses are uncorrelated, the two-sample t-test estimator achieves the same precision as ANCOVA II but is inefficient otherwise, while the paired t-test is equivalent to ANCOVA II only if the difference between follow-up and baseline is uncorrelated with baseline within each treatment.

As the ANCOVA approaches are derived from a supposed linear relationship between baseline and follow-up, some practitioners are reluctant to use them; asymptotic equivalence of the GEE estimator indicates it involves the same considerations. Yang and Tsiatis (2001) showed that consistency and asymptotic normality hold even if linearity is violated. However, as their study restricted attention to such "linear" estimators, it did not address whether or how it might be possible to improve upon these approaches under deviations from linearity and without limiting distributional assumptions.

Such deviations are commonplace, as illustrated by data from AIDS Clinical Trials Group (ACTG) protocol 175 involving 2467 HIV-infected subjects randomized originally to four treatment groups, zidovudine (ZDV) alone, ZDV plus didanosine (ddI), ZDV plus zalcitabine (ddC), or ddI alone in approximately equal numbers (Hammer et al., 1996). Analysis of the primary endpoint of time to progression to AIDS or

death showed ZDV to be inferior to the other three therapies, which showed no differences. Figure 1.1 plots CD4 counts at 20 ± 5 weeks, a follow-up measure that reflects early response to treatment subsequent to the often-observed initial rise in CD4 (e.g. Tsiatis, DeGruttola, and Wulfsohn, 1995), versus baseline CD4 for the ZDV-only ("control") and other therapies combined ("treatment") groups and suggests a possible departure from a straight-line relationship. Figure 1.2, which plots CD4 counts at 96 ± 5 weeks, shows similar behavior. Indeed, nonlinear relationships are a routine feature of biological phenomena, as is nonnormality; histograms of CD4 counts at baseline and follow-up for each group (not shown) exhibit the usual asymmetry that motivates the standard analysis on the log, fourth-, or cube-root scale.

A further complication facing the data analyst, particularly in lengthy studies, is that of missing follow-up response for some subjects, the proportion of whom may be nontrivial. A typical approach of this problem is to undertake a "complete-case" analysis, disregarding data from subjects with missing follow-up and using one of the above techniques; with GEE, data on all subjects may be used. However, as is well known, unless these data are missing completely at random (MCAR) (Rubin 1976), these strategies may yield biased inference on β. Often, baseline demographic and physiologic characteristics are collected on each participant; moreover, during the intervening period from baseline to follow-up, additional covariate information, including intermediate measures of the response, may be obtained. Missingness at follow-up is often associated with baseline response and baseline and intermediate covariates, a relationship that may be differential by intervention. Under such circumstances,

the assumption that follow-up is missing at random (MAR) (Rubin 1976), associated only with these observable quantities and not the unobserved response, may be reasonable. Even if such a MAR mechanism may be postulated, valid methods to take appropriate account of missingness in pretest-posttest analysis have not been widely applied by practitioners, and, rather, *ad hoc* approaches such as complete-case analysis have been commonplace.

Semiparametric models, written in terms of a finite-dimensional parameter of interest and an unspecified infinite-dimensional component, have gained considerable popularity as they embody less restrictions (and thus fewer potentially incorrect assumptions) than fully parametric models. In a landmark paper, Robins, Rotnitzky, and Zhao (1994) derived an asymptotic theory of inference for general semiparametric models with data MAR. Generically, for parameter β in a statistical model, an estimator $\widehat{\beta}$ based on iid observations W_i, $i = 1, \ldots, n$, is asymptotically linear with influence function $\varphi(W)$ if $n^{1/2}(\widehat{\beta}-\beta) = n^{-1/2}\sum_{i=1}^{n}\varphi(W_i)+o_p(1)$ and $E\{\varphi(W)\} = 0$, $E\{\varphi^T(W)\varphi(W)\} < \infty$; regular, asymptotically linear (RAL) estimators (Newey 1990) $\widehat{\beta}$ are consistent and asymptotically normal under weak conditions with asymptotic variance $E\{\varphi^T(W)\varphi(W)\}$. Thus, by identifying all influence functions for a particular model, consistent estimators may be deduced. Robins et al. (1994) not only characterized the class of influence functions for all RAL estimators for the parametric component of general semiparametric models with data MAR, but also identified the efficient member of the class with smallest variance.

In light of these developments, rather than study the pretest-posttest problem with

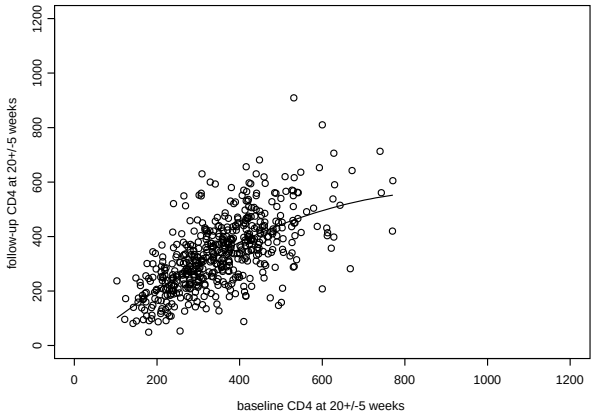
MAR data under specific distributional assumptions or adapt "popular" estimators such as ANCOVA to handle MAR on a case-by-case basis, a more broadly-applicable strategy is to take a semiparametric perspective and use this theory to elucidate a unified framework for pretest-posttest analysis. Interestingly, despite the ubiquity of the pretest-posttest study in numerous fields and the simplicity of the model when no data are missing, to our knowledge, explicit application of this pioneering theory to pretest-posttest inference with data MAR with an eye toward development of practical estimators on its basis and associated detailed study and illustration of performance of resulting techniques has not been reported.

A main objective of this dissertation is to develop practical strategies for estimation of β in a semiparametric pretest-posttest model from the perspective of the Robins et al. (1994) theory by indicating the class of influence functions for all RAL estimators for β when follow-up response is MAR, including the most efficient. We first apply this theory in this setting, using it to identify the class of all influence functions when no follow-up data are missing by casting the full-data situation as an "artificial" missing data problem by defining counterfactual follow-up responses (e.g. Holland 1986). A key finding is that the resulting estimators exploit the relationship of baseline response and covariates to follow-up response in a way that leads to substantial efficiency gains over "popular" approaches such as ANCOVA or the paired t-test, which are revealed to yield inefficient members of the class of consistent estimators for β.

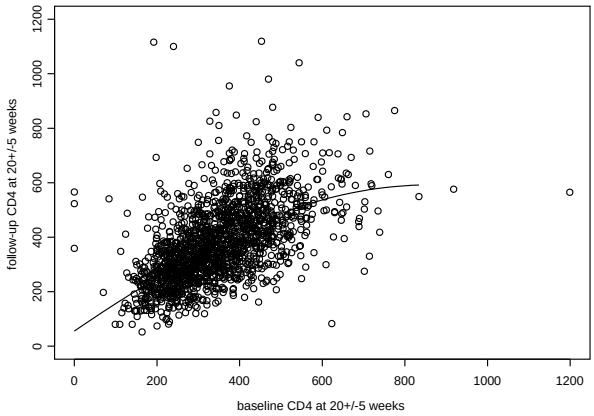
Armed with this developments, we explicate how relationships among responses

and covariates play a role in enhancing precision, demonstrate how the theory leads to practical estimators, and study the performance of such approaches.

In Chapter 2, we first introduce the model and derive the semiparametric class of full-data influence functions using counterfactuals. Then in Section 2.2, we describe how the class of all observed data influence functions under MAR follows from the Robins et al. (1994) theory. Chapter 3 presents strategies for constructing estimators, and performance is demonstrated via simulation in Chapter 4 and by application to data from an HIV clinical trial in Chapter 5, for both full data and observed data cases. As in any missing-data context, plausibility of the MAR assumption for follow-up response is critical and may be best justified with availability of rich baseline and intervening information.

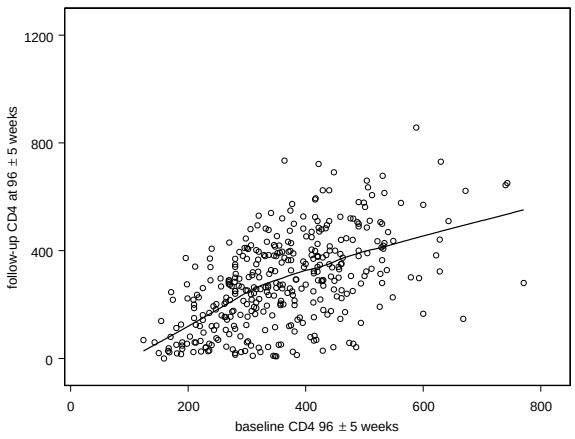


(a)

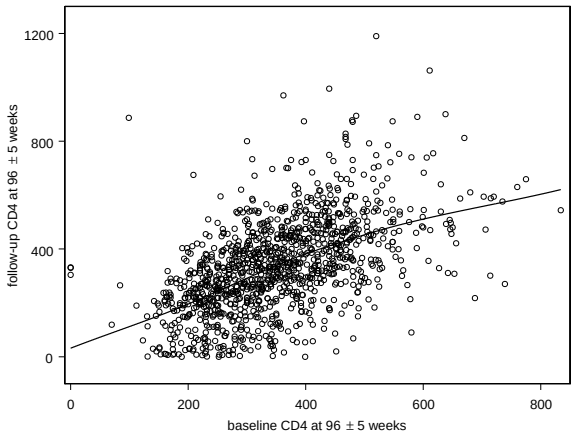


(b)

Figure 1.1: ACTG 175: CD4 counts after 20 ± 5 weeks vs. baseline CD4 counts for patients randomly assigned to **a.** ZDV alone ("control") and **b.** the combination of ZDV+ddI, ZDV+ddC, or ddI alone ("treatment"). The solid lines were obtained using the Splus function `loess()` (Cleveland, Gross, and Shyu, 1993); in **b.**, use of different span values and deletion of the apparent "high leverage" points with the largest baseline CD4 all lead to similar curvilinear fits.



(a)



(b)

Figure 1.2: ACTG 175: CD4 counts after 96 ± 5 weeks vs. baseline CD4 counts for patients randomly assigned to **a.** ZDV alone and **b.** the combination of ZDV+ddI, ZDV+ddC, or ddI alone. The solid lines were obtained using the Splus function `loess()` (Cleveland, 1979); in **b.** use of default span values.

Chapter 2

MODEL

2.1 Full data model

In this section, assume no missing follow-up and that each subject $i = 1, \ldots, n$ is randomized to treatment with known probability δ, so $Z_i = 0$ or 1 as i is assigned to control or treatment, respectively. Let Y_{1i} and Y_{2i} be i's observed baseline and follow-up responses, leading to observed data for i (Y_{1i}, Y_{2i}, Z_i); the subscript i is suppressed when no ambiguity will result.

We develop the model by conceptualizing the situation in terms of counterfactuals or potential outcomes, a key device in the study of causal inference (e.g., Holland, 1986), and then expressing the observable data in terms of these quantities. The variables Y_1 and Z represent phenomena prior to treatment action, while Y_2 is a post-treatment characteristic. Thus, let $Y_2^{(0)}, Y_2^{(1)}$ be the follow-up responses a subject potentially would exhibit if assigned to control and treatment, respectively. The full

set of counterfactual random variables $(Y_2^{(0)}, Y_2^{(1)})$ is obviously not observable for any subject but rather represents what potentially might occur at follow-up under both treatments, including that "counter to the fact" of what might actually be assigned in the trial. We place no restrictions on the joint distribution of the counterfactuals and Y_1, such as equal variance or independence, and define $\mu_1 = E(Y_1)$, $\sigma_{11} = \text{var}(Y_1)$; for $c = 0, 1$, $\mu_2^{(c)} = E(Y_2^{(c)})$, $\sigma_{22}^{(c)} = \text{var}(Y_2^{(c)})$; and $\sigma_{12}^{(c)} = \text{cov}(Y_1, Y_2^{(c)})$. Thus, e.g., $\mu_2^{(0)} = E(Y_2^{(0)})$ denotes mean follow-up response if all subjects in the population were assigned to control. It is natural to assume that the observed follow-up response under the subject's actual, assigned treatment corresponds to what potentially would be seen if the subject were assigned to that treatment; i.e., $Y_2 = Y_2^{(0)}(1 - Z) + Y_2^{(1)}Z$. The observed assignment Z is made at random, so without regard to baseline status or prognosis; thus, assume Z is independent of $(Y_1, Y_2^{(0)}, Y_2^{(1)})$. As usual, we assume $(Y_{1i}, Y_{2i}^{(0)}, Y_{2i}^{(1)}, Z_i)$ and hence (Y_{1i}, Y_{2i}, Z_i) are independent and identically distributed (i.i.d.) across i.

Interest focuses on the difference in population mean follow-up response, which, under the usual causal inference perspective (Holland, 1986), may be thought of as the difference in means if all subjects in the population were assigned to control or treatment, respectively; i.e., $\beta = \mu_2^{(1)} - \mu_2^{(0)} = E(Y_2^{(1)}) - E(Y_2^{(0)})$. Under our assumptions, in fact $\beta = E(Y_2|Z = 1) - E(Y_2|Z = 0)$, the usual expression for the difference of interest in a randomized trial; and $E(Y_2|Z) = \mu_2 + \beta Z$ and $E(Y_1|Z) = \mu_1$, writing $\mu_2 = \mu_2^{(0)} = E(Y_2^{(0)})$ for brevity.

The advantage of this framework and expression of β in terms of counterfactual

means is that it reveals a structure analogous to that in missing data problems. As we are interested in the difference of the two marginal quantities $\mu_2^{(1)}$ and $\mu_2^{(0)}$, we may view estimation of each mean separately, without reference to the joint distribution of $Y_2^{(0)}, Y_2^{(1)}$; thus, if we identify estimators for $\mu_2^{(1)}$ and $\mu_2^{(0)}$ that are "optimal" in some sense, the "optimal" estimator for β may be obtained as their difference. Accordingly, focus on $\mu_2^{(1)}$; considerations for $\mu_2^{(0)}$ are similar. If we could observe the "full data" $(Y_1, Y_2^{(1)}, Z)$ for all n subjects, then we would would estimate $\mu_2^{(1)}$ by the sample mean $n^{-1}\sum_{i=1}^n Y_{2i}^{(1)}$. However, we only observe $Y_2^{(1)}$ for subjects with observed assignment $Z = 1$, so that $Y_2^{(1)}$ is "missing" for subjects with $Z = 0$; i.e., we only observe $(Y_1, ZY_2^{(1)}, Z)$, where $Y_2^{(1)}$ is observed with probability $P(Z = 1|Y_1, Y_2^{(1)}) = P(Z = 1) = \delta$, so that $Y_2^{(1)}$ is "missing completely at random" (MCAR; Rubin, 1976). Thus, the two-sample t-test estimator based on observed sample averages,

$$n_1^{-1}\sum_{i=1}^n Z_i Y_{2i} - n_0^{-1}\sum_{i=1}^n (1 - Z_i)Y_{2i}, \quad n_1 = \sum_{i=1}^n Z_i, \; n_0 = \sum_{i=1}^n (1 - Z_i) \tag{2.1.1}$$

may be regarded as a "complete case" estimator for β. As such, while it may be unbiased for β under MCAR, it is likely inefficient, as it takes no account of observations on Y_1. This suggests that a more refined approach to missing data problems may lead to improved estimators that exploit information in Y_1 and the interrelationships among observed variables.

Robins et al. (1994) developed a general large-sample theory of estimation in semiparametric models where data are missing at random (so including MCAR). They described the class of all (regular, excluding "pathological" cases) estimators under these conditions by characterizing the form of the influence functions of all members

of the class of asymptotically linear estimators (e.g., Newey, 1990), which are consistent and asymptotically normal under general conditions. An estimator $\widehat{\beta}$ for β is asymptotically linear with influence function $\mathcal{I}$ if $n^{1/2}(\widehat{\beta}-\beta) = n^{-1/2}\sum_{i=1}^{n}\mathcal{I}_i + o_p(1)$ and $E(\mathcal{I}) = 0$, $E(\mathcal{I}^T\mathcal{I}) < \infty$, and the asymptotic variance of $\widehat{\beta}$ is the variance of the influence function. Robins et al. (1994) identified the most efficient regular, asymptotically linear (RAL) estimator, namely, that whose influence function has smallest variance. We apply this general theory to our model to characterize all estimators for β through their influence functions. This allows us not only to demonstrate that the two-sample t-test, paired t-test, and ANCOVA I and II estimators are inefficient members of this class but also to elucidate the form of the most efficient estimator for β. If for each $c = 0, 1$ we were able to observe the "full data" $(Y_1, Y_2^{(c)}, Z)$ for all subjects, we would estimate $\mu_2^{(c)}$ by the sample mean $n^{-1}\sum_{i=1}^{n} Y_{2i}^{(c)}$, with influence function $\varphi^{(c)}(Y_2^{(c)}) = Y_2^{(c)} - \mu_2^{(c)}$. Because $\mu_2^{(c)}$, $c = 0, 1$, is an explicit function of the distribution of $Y_2^{(c)}$, which we take to be unrestricted, $\varphi^{(0)}$ and $\varphi^{(1)}$ are the only such "full data" influence functions for estimators for $\mu_2^{(0)}$ and $\mu_2^{(1)}$ (Newey, 1990). Of course, we only observe (Y_1, Y_2, Z) for each subject; thus, we require estimators that may be expressed in terms of these quantities. By the analogy to missing data problems, it follows from the theory of Robins et al. (1994) that all RAL estimators for β based on the observed data have influence function of form

$$\left\{\frac{Z\varphi^{(1)}(Y_2)}{\delta} + \frac{(Z-\delta)}{\delta}h^{(1)}(Y_1)\right\} - \left[\frac{(1-Z)\varphi^{(0)}(Y_2)}{1-\delta} + \frac{\{(1-Z)-(1-\delta)\}}{1-\delta}h^{(0)}(Y_1)\right], \tag{2.1.2}$$

where $h^{(0)}, h^{(1)}$ are arbitrary functions such that $\text{var}\{h^{(c)}(Y_1)\} < \infty$, $c = 0, 1$; (2.1.2) is the difference of the forms of all observed data influence functions for estimators for $\mu_2^{(1)}$ and $\mu_2^{(0)}$. For arbitrary h with $\text{var}\{h(Y_1)\} < \infty$, we may rewrite (2.1.2) as

$$Z(Y_2 - \mu_2 - \beta)/\delta - (1 - Z)(Y_2 - \mu_2)/(1 - \delta) + (Z - \delta)h(Y_1), \tag{2.1.3}$$

Thus, (2.1.3) characterizes all consistent estimators for β, and we expect that the influence functions for the "popular" estimators in the next section may be represented in this form.

Popular Estimators

The two-sample t-test estimator for β is given in (2.1.1). The paired t-test estimator is $\overline{D}_1 - \overline{D}_0$, where $\overline{D}_1 = n_1^{-1} \sum_{i=1}^n Z_i(Y_{2i} - Y_{1i})$ and $\overline{D}_0 = n_0^{-1} \sum_{i=1}^n (1 - Z_i)(Y_{2i} - Y_{1i})$. As in Yang and Tsiatis (2001), the ANCOVA I estimator for β is obtained by ordinary least squares (OLS) regression of Y_2 on (Y_1, Z), and the ANCOVA II estimator is obtained by OLS regression of $(Y_2 - \overline{Y}_2)$ on $\{(Y_1 - \overline{Y}_1), (Z_i - \overline{Z}), (Y_1 - \overline{Y}_1)(Z_i - \overline{Z})\}^T$. It is straightforward to show that all of these estimators have influence functions of the form

$$\frac{Z(Y_2 - \mu_2 - \beta)}{\delta} - \frac{(1 - Z)(Y_2 - \mu_2)}{1 - \delta} + (Z - \delta)(Y_1 - \mu_1)\eta, \tag{2.1.4}$$

where $\eta = 0,\ -1/\{\delta(1 - \delta)\},\ -\{\delta\sigma_{12}^{(1)} + (1 - \delta)\sigma_{12}^{(0)}\}/\{\sigma_{11}\delta(1 - \delta)\}$, and $-\{(1 - \delta)\sigma_{12}^{(1)} + \delta\sigma_{12}^{(0)}\}/\{\sigma_{11}\delta(1 - \delta)\}$ for the two-sample t-test, paired t-test, ANCOVA I, and ANCOVA II estimators, respectively. Thus, from (2.1.4), these estimators are all in class (2.1.3) with $h(Y_1) = \eta(Y_1 - \mu_1)$ and hence are consistent and asymptotically normal.

The observations of Yang and Tsiatis (2001) are immediate: If $\delta = 0.5$, η is identical for ANCOVA I and II and these estimators are asymptotically equivalent. If $\sigma_{12}^{(c)} = 0$, $c = 0, 1$, i.e., uncorrelated baseline and follow-up, $\eta = 0$ for ANCOVA I/II and both are asymptotically equivalent to the two-sample t-test. Finally, if $Y_2^{(c)} - Y_1$ is uncorrelated with Y_1 so $\sigma_{12}^{(c)} = \sigma_{11}$, $c = 0, 1$, the paired t-test is equivalent to ANCOVA I/II. Interestingly, while η values for ANCOVA I/II both involve a "weighted average" of $\sigma_{12}^{(0)}$ and $\sigma_{12}^{(1)}$, the "weighting" for the latter seems counterintuitive in that $(1-\delta) = P(Z = 0)$ and $\delta = P(Z = 1)$ are the coefficients of the covariances for treatments 1 and 0, whereas one might expect the reverse.

Not only are all the "popular" estimators in class (2.1.3), they belong to the subclass with h a linear function of Y_1, suggesting there is room for improvement via more general h.

We close this section by sketching steps to identify the most efficient influence function φ^{eff} in class (2.1.3). Following the theory of Robins et al. (1994), the set of all functions of (Y_2, Y_1, Z) of the form $(Z-\delta)h(Y_1)$, $E\{h^2(Y_1)\} < \infty$, is a linear subspace of the Hilbert space of all mean-zero functions $\varphi(Y_1, Y_1, Z)$ with $E\{\varphi^2(Y_1, Y_1, Z)\} < \infty$ with covariance inner product. Denoting this subspace as Λ_2, the most efficient estimator for β is that with influence function

$$\left\{ \frac{Z(Y_2 - \mu_2 - \beta)}{\delta} - \frac{(1-Z)(Y_2 - \mu_2)}{(1-\delta)} \right\} - \Pi \left\{ \frac{Z(Y_2 - \mu_2 - \beta)}{\delta} - \frac{(1-Z)(Y_2 - \mu_2)}{(1-\delta)} \middle| \Lambda_2 \right\}, \tag{2.1.5}$$

where $\Pi(\cdot\,|\Lambda_2)$ is the projection of the argument onto Λ_2. Because projection is a linear operation, the projection may be found as the difference of the projections of the

components of the first term in (2.1.5). To find $\Pi\{Z(Y_2-\mu_2-\beta)/\delta|\,\Lambda_2\}$, we must find $h^{(1)}(Y_1)$ such that $E\left[\{Z(Y_2-\mu_2-\beta)/\delta\ -\ (Z-\delta)h^{(1)}(Y_1)\}\,(Z-\delta)h(Y_1)\right]=0$ for all $h(Y_1)$; i.e., we require that $E\left[\{Z(Y_2-\mu_2-\beta)/\delta\ -\ (Z-\delta)h^{(1)}(Y_1)\}\,(Z-\delta)|\,Y_1\right]=0$ a.s. This may be written equivalently as $E\{Z(Y_2-\mu_2-\beta)(Z-\delta)/\delta|\,Y_1\}=h^{(1)}(Y_1)\,E\{(Z-\delta)^2|\,Y_1\}$ a.s. Using independence of Z and Y_1, the left-hand side of this expression equals $\{E(Y_2^{(1)}|\,Y_1)-\mu_2-\beta\}(1-\delta)$ and $E\{(Z-\delta)^2|\,Y_1\}=\delta(1-\delta)$, so that $h^{(1)}(Y_1)=\{E(Y_2^{(1)}|\,Y_1)-\mu_2-\beta\}/\delta$, which yields $\Pi\{Z(Y_2-\mu_2-\beta)/\delta|\Lambda_2\}=(Z-\delta)\{E(Y_2^{(1)}|Y_1)-\mu_2-\beta\}/\delta$. Similarly, $\Pi\{Z(Y_2-\mu_2)/(1-\delta)|\Lambda_2\}=(Z-\delta)\{E(Y_2^{(0)}|Y_1)-\mu_2\}/(1-\delta)$. Substituting into (2.1.5) yields the result:

$$\left[\frac{Z(Y_2-\mu_2-\beta)}{\delta}+\frac{(Z-\delta)\{E(Y_2|X_0,Y_1,Z=1)-\mu_2-\beta\}}{\delta}\right]$$
$$-\quad\left[\frac{(1-Z)(Y_2-\mu_2)}{1-\delta}+\frac{(Z-\delta)\{E(Y_2|X_0,Y_1,Z=0)-\mu_2\}}{1-\delta}\right]. \qquad (2.1.6)$$

In Chapter 3 we discuss approaches to deducing estimators for β from this result when full data are available that offer dramatic improvements in efficiency over "popular" methods.

2.2 Observed data influence functions

Now suppose Y_2 is missing for some subjects, with all other variables observed, and define $R=0$ or 1 as Y_2 is missing or observed. Then the data observed for subject i may be represented as $O_i=(X_{0i},Y_{1i},X_{1i},R_i,R_iY_{2i},Z_i)$. We formalize the assumption that Y_2 is MAR as $P(R=1|Y_1,X_0,X_1,Y_2,Z)=P(R=1|Y_1,X_0,X_1,Z)=\pi(Y_1,X_0,X_1,Z)\geq\epsilon>0$, reflecting the reasonable view for a pretest-posttest trial

that there is a positive probability of observing Y_2 for any subject. Here, dependence of the missingness mechanism on all data is taken not to involve the unobserved Y_2 but may be associated with baseline and intermediate characteristics and be differential by intervention, the latter highlighted by the equivalent representation $\pi(Y_1, X_0, X_1, Z) = Z\pi^{(1)}(Y_1, X_0, X_1) + (1 - Z)\pi^{(0)}(Y_1, X_0, X_1)$ for $\pi^{(c)}(Y_1, X_0, X_1) = \pi(Y_1, X_0, X_1, c) \geq \epsilon > 0,\ c = 0, 1$.

As noted in Chapter 1, it is common under these conditions to conduct a complete-case analysis. E.g., using the two-sample t-test, estimate β by the difference in sample means based only on data for subjects with Y_2 observed, yielding $\widehat{\beta} = \sum_{i=1}^{n} R_i Z_i Y_{2i}/n_{R1} - \sum_{i=1}^{n} R_i(1 - Z_i)Y_{2i}/n_{R0}$, where $n_{Rc} = \sum_{i=1}^{n} R_i I(Z_i = c),\ c = 0, 1$. It is easy to verify that this estimator and, indeed, those for each mean, are not consistent for β and $\mu_2^{(c)},\ c = 0, 1$, respectively. A simple remedy is to incorporate "inverse weighting" of the complete cases (IWCC; e.g. Horvitz and Thompson 1952). For example, noting the complete-case estimator for $\mu_2^{(1)}$ solves $\sum_{i=1}^{n} R_i Z_i(Y_{2i} - \mu^{(1)}) = 0$, weight each contribution by the inverse of the probability of seeing a complete case; i.e., solve $\sum_{i=1}^{n} R_i Z_i(Y_{2i} - \mu^{(1)})/\pi_i^{(1)}(X_{0i}, Y_{1i}, X_{1i}) = 0$, yielding the estimator $\overline{Y}_2^{(1)} = \{\sum_{i=1}^{n} R_i Z_i Y_{2i}/\pi^{(1)}(X_{0i}, Y_{1i}, X_{1i})\}/n_{RZ1},\ n_{RZ1} = \sum_{i=1}^{n} R_i Z_i/\pi^{(1)}(X_{0i}, Y_{1i}, X_{1i})$, and analogously for $\mu_2^{(0)}$. It is straightforward to show that $\overline{Y}_2^{(c)}$ are consistent for $\mu_2^{(c)}$, $c = 0, 1$, and to find the associated influence functions given by

$$\frac{RZ(Y_2 - \mu_2^{(1)})}{\delta\pi^{(1)}(X_0, Y_1, X_1)} \quad \text{and} \quad \frac{R(1 - Z)(Y_2 - \mu_2^{(0)})}{(1 - \delta)\pi^{(1)}(X_0, Y_1, X_1)}; \tag{2.2.7}$$

each has the form of the corresponding full-data influence function weighted by $1/\pi^{(1)}$

for the complete cases only. IWCC may be similarly applied to any RAL estimator with influence function in class (2.1.3), including "popular" ones, to yield consistent inference. However, although simple IWCC leads to consistency, methods with greater efficiency are possible.

The pioneering advance of Robins et al. (1994) was to derive for a general semiparametric model the class of all influence functions for a parametric component based on the data observed under complex forms of MAR, where different subsets of the full data may be missing in different ways, and to characterize the efficient influence function. The development follows geometric principles, complicated by the need to distinguish between the ideal, full data, denoted by V here, and the data O observed under MAR. Now, the Hilbert space $\mathcal{H}$ in which influence functions are elements is that of all mean-zero, finite-variance random functions $h(O)$, with inner product $E(h_1^T h_2)$. The key is to identify the corresponding nuisance tangent space, and hence representation of influence functions, which is a considerably more complex and delicate enterprise than that for the full-data problem sketched in Section 2.1. The theory reveals, perhaps not unexpectedly, that there is a relationship between influence functions based on the full and observed data. In particular, when the function describing the probability that full data are observed, $\pi(O^*)$, is known, as we assume for now, which under MAR depends only on the subset of V that is always observed, denoted O^*, then if $\varphi^F(V)$ is any full-data influence function, Robins et al. showed that all observed-data influence functions under MAR have the form $R\varphi^F(V)/\pi(O^*) + g(O)$, where $g(O)$ is an arbitrary function of the observed data sat-

isfying $E\{g(O)|V\} = 0$. In cases like that here, where a single subset of V is either missing or not for all subjects, this becomes

$$\frac{R\varphi^F(V)}{\pi(O^*)} + \frac{R - \pi(O^*)}{\pi(O^*)} g(O^*) \tag{2.2.8}$$

where now $g(O^*)$ is an arbitrary function of the data always observed. In (2.2.8), note that the first term has the form of an IWCC full-data influence function; the second term, which has mean zero, depending only on data observed for all subjects, "augments" (e.g. Robins 1999) the first, which leads to increased efficiency provided that g is chosen judiciously.

In the special case of the pretest-posttest problem, focusing on estimation of the treatment mean $\mu_2^{(1)} = \mu_2 + \beta$, with $O^* = (X_0, Y_1, X_1, Z)$, (2.1.2) and (2.2.8) imply that the class of all observed-data influence functions when Y_2 is MAR is

$$\frac{R\{Z(Y_2 - \mu_2^{(1)}) - (Z - \delta)h^{(1)}(Y_1, X_0)\}}{\delta\pi(Y_1, X_0, X_1, Z)} - \frac{R - \pi(Y_1, X_0, X_1, Z)}{\pi(Y_1, X_0, X_1, Z)} g^{(1)}(Y_1, X_0, X_1, Z), \tag{2.2.9}$$

for arbitrary $h^{(1)}$ and $g^{(1)}$ such that $\text{var}\{h^{(1)}(X_0, Y_1)\} < \infty$ and $\text{var}\{g^{(1)}(Y_1, X_0, X_1, Z)\} < \infty$.

Robins et al. (1994) provide a general mechanism for deducing the optimal choices of $h^{(1)eff}$ and $g^{(1)eff}$ leading to the efficient influence function. For special cases like (2.2.8) and (2.2.9), it is straightforward and instructive to identify these choices and to gain insight into how "augmentation" increases efficiency over IWCC estimators directly via geometric arguments. To this end, it proves convenient to define $g^{(1)'}(Y_1, X_0, X_1, Z) = (Z - \delta)h^{(1)}(Y_1, X_0) + \delta g^{(1)}(Y_1, X_0, X_1, Z)$. We may write (2.2.9)

equivalently as

$$\frac{RZ(Y_2-\mu_2^{(1)})}{\delta\pi(Y_1,X_0,X_1,Z)}-\frac{(Z-\delta)}{\delta}h^{(1)}(Y_1,X_0)-\frac{R-\pi(Y_1,X_0,X_1,Z)}{\delta\pi(Y_1,X_0,X_1,Z)}g^{(1)'}(Y_1,X_0,X_1,Z); \tag{2.2.10}$$

there is a one-to-one correspondence between (2.2.9) and (2.2.10). It is straightforward to show that the second and third terms in (2.2.10) are uncorrelated and thus define orthogonal subspaces of mean-zero functions in $\mathcal{H}$. Writing (2.2.10) as $A-B_1-B_2$, it is easy to deduce that minimizing the variance under these conditions is equivalent to minimizing the variances of $A-B_1$ and $A-B_2$ separately, which may be viewed as finding the separate projections of A onto the spaces defined by B_1 and B_2. Thus, as in the argument for the efficient full-data influence function at the end of Section 2.1, dividing by known δ, we wish to find $h^{eff(1)}$ and $g^{eff(1)'}$ such that $E[\,\{RZ(Y_2-\mu_2^{(1)})/\pi(Y_1,X_0,X_1,Z)-(Z-\delta)h^{eff(1)}(X_0,Y_1)\}(Z-\delta)h^{(1)}(X_0,Y_1)\,]=0$ and $E\Big(\,[RZ(Y_2-\mu_2^{(1)})/\pi(Y_1,X_0,X_1,Z)-g^{eff(1)'}(Y_1,X_0,X_1,Z)\{R-\pi(Y_1,X_0,X_1,Z)\}/\pi(Y_1,X_0,X_1,Z)]g^{(1)'}(Y_1,X_0,X_1,Z)\{R-\pi(Y_1,X_0,X_1,Z)\}/\pi(Y_1,X_0,X_1,Z)\,\Big)=0$ for all $h^{(1)}$ and $g^{1)'}$. By a conditioning argument as in Section 2.1, it may be deduced that $h^{eff(1)}(X_0,Y_1)=E(Y_2|X_0,Y_1,Z=1)-\mu_2^{(1)}$ and $g^{eff(1)'}(X_0,Y_1)=Z\{E(Y_2|X_0,Y_1,X_1,Z)-\mu_2^{(1)}\}=Z\{E(Y_2|X_0,Y_1,X_1,Z=1)-\mu^{(1)_2}\}$. That is,

$$\begin{aligned}&\frac{RZ(Y_2-\mu_2-\beta)}{\delta\pi(Y_1,X_0,X_1,Z=1)}-\frac{(Z-\delta)}{\delta}\{E(Y_2|Y_1,X_0,Z=1)-\mu_2^{(1)}\}\\ -\;&\frac{R-\pi(Y_1,X_0,X_1,Z=1)}{\delta\pi(Y_1,X_0,X_1,Z=1)}E(Y_2|Y_1,X_0,X_1,Z=1)-\mu_2^{(1)};\end{aligned} \tag{2.2.11}$$

Note that $g^{eff(1)'}$ does not depend on $h^{eff(1)}$, and $h^{eff(1)}$ is identical to the optimal choice for the full-data case. These features *need not* hold for general semiparametric

models; here they are a consequence the simple structure of the pretest-posttest problem.

From the form of $g^{eff(1)'}$ and the representation of π, for the purpose of finding estimators "close to" that with efficient form, discussed in the next section, it thus suffices to restrict attention to the subclass of (2.2.10) with $g^{(1)'}(X_0, Y_1, X_1, Z) = Zq^{(1)}(X_0, Y_1, X_1)$ for arbitrary square-integrable $q^{(1)}$ given by

$$\begin{aligned}\psi(X_0, Y_1, X_1, R, RY_2, Z) &= \frac{RZ(Y_2 - \mu_2^{(1)})}{\delta\pi^{(1)}(X_0, Y_1, X_1)} - \frac{(Z-\delta)}{\delta}h^{(1)}(X_0, Y_1) \\ &- \frac{\{R - \pi^{(1)}(X_0, Y_1, X_1)\}Z}{\delta\pi^{(1)}(X_0, Y_1, X_1)}q^{(1)}(X_0, Y_1, X_1); \quad (2.2.12)\end{aligned}$$

(2.2.12) includes the optimal $g^{(1)'}$ but rules out choices that cannot have the efficient form.

The foregoing results take π and hence $\pi^{(1)}$ to be known, which is unlikely in practice unless Y_2 is missing purposefully by design for some subjects depending on a subject's baseline and intermediate information. In practice, this is usually addressed by positing a parametric model for $\pi^{(1)}$ in terms of a $(s \times 1)$ parameter γ. Thus, to exploit (2.2.12) to derive consistent estimators when $\pi^{(1)}$ is not known, intuition suggests that such a model be correctly specified, although we discuss this further below. Hence, write $\pi^{(1)}(X_0, Y_1, X_1; \gamma)$; e.g., for definiteness, consider a logistic regression model $\pi^{(1)}(X_0, Y_1, X_1; \gamma) = \exp\{\gamma^T d(X_0, Y_1, X_1)\}/[1 + \exp\{\gamma^T d(X_0, Y_1, X_1)\}]$, where $d(X_0, Y_1, X_1)$ is a vector of functions of its argument. This introduces an additional parametric component in the semiparametric model, and implementation would require that γ be estimated from the data (X_0, Y_1, X_1, R, Z) and substituted in estimators derived from (2.2.12). It may be shown using the Robins et al. (1994) theory

in this case that, as long as an efficient procedure is used to estimate γ, the class of influence functions containing the optimal $g^{(1)\prime}$ corresponding to (2.2.12) is

$$\psi(X_0, Y_1, X_1, R, RY_2, Z) + d^T(X_0, Y_1, X_1)A_{(1)}^{-1}(b_{q(1)} - b_{(1)})\frac{\{R - \pi^{(1)}(X_0, Y_1, X_1)\}Z}{\delta}, \tag{2.2.13}$$

where $b_{(1)} = E[(Y_2 - \mu_2^{(1)})\{1 - \pi^{(1)}(X_0, Y_1, X_1)\}d(X_0, Y_1, X_1)|Z = 1]$, $b_{q(1)} = E[q^{(1)}(X_0, Y_1, X_1)\{1 - \pi^{(1)}(X_0, Y_1, X_1)\}d(X_0, Y_1, X_1)|Z = 1]$ and $A_{(1)} = E[\pi^{(1)}(X_0, Y_1, X_1)\{1 - \pi^{(1)}(X_0, Y_1, X_1)\}d(X_0, Y_1, X_1)d^T(X_0, Y_1, X_1)|Z = 1]$; and $\pi^{(1)}(X_0, Y_1, X_1)$ is evaluated at the true value of γ, which is suppressed here and in the sequel unless otherwise specified.

Several key results with implications for practice follow from the Robins et al. (1994) theory. Estimators for $\mu_2^{(1)}$ with influence functions in class (2.2.13) may be found by finding estimators with influence functions in class (2.2.12) (so for γ known) and substituting the maximum likelihood (ML) estimator for γ. Modeling the mechanism separately for $Z = 0, 1$ rather than jointly as $\pi(X_0, Y_1, X_1, Z)$, although restricting the class of missingness models, does not restrict the class of efficient estimators; in fact, if the true relationship follows a parametric model $\pi(X_0, Y_1, X_1, Z; \gamma)$, inducing models $\pi^{(c)}(Y_1, X_0, X_1; \gamma)$, $c = 0, 1$, estimating γ separately by ML for $Z = 0, 1$ will lead to a more efficient estimator for $\mu_2^{(1)}$ than that found by estimating γ jointly. Thus, we take γ to be estimated separately for $Z = 0, 1$.

In the case where $q^{(1)}(Y_1, X_0, X_1)$ has the efficient form $E(Y_2|X_0, Y_1, X_1, Z = 1) - \mu_2^{(1)}$, $b_{(1)} = b_{q(1)}$, and hence, even if γ is estimated, the last term in (2.2.13)

is identically equal to zero, but this will not necessarily be true otherwise. This reflects the general result shown by Robins et al. (1994) that an estimator derived from the efficient influence function will have the same properties whether γ is known (so known missingness mechanism) or estimated.

In fact, the theory also implies the seemingly counterintuitive result that, even if γ is known, it is possible to gain efficiency by estimating it anyway; i.e., for a particular choice of $h^{(1)}$ and $q^{(1)}$, the variance of an influence function of form (2.2.13) is smaller than that of (2.2.12). Geometrically, this is because (2.2.13) is the residual found from projection of $\psi(X_0, Y_1, X_1, R, RY_2, Z)$ onto the linear subspace of $\mathcal{H}$ spanned by the score for γ when γ is estimated from data with $Z = 1$ only, $S_\gamma(X_0, Y_1, X_1, Z; \gamma_0) = d(X_0, Y_1, X_1)\{R - \pi^{(1)}(X_0, Y_1, X_1; \gamma_0)\}Z$, given by $\{BS_\gamma(X_0, Y_1, X_1, Z)$ for all $(p \times s)$ matrices $B\}$, where γ_0 is the true value of γ. To see this, consider the simple, special case of the IWCC in (2.2.7), so with $h^{(1)} \equiv q^{(1)} \equiv 0$. Then $b_{q(1)} = 0$, and the projection of S_γ onto this space of form $B_0 S_\gamma(X_0, Y_1, X_1, Z, \gamma_0)$ must satisfy $E\Big([RZ(Y_2 - \mu_2^{(1)})/\{\delta\pi^{(1)}(X_0, Y_1, X_1; \gamma_0)\} - B_0 S_\gamma(X_0, Y_1, X_1, Z, \gamma_0)](BS_\gamma(X_0, Y_1, X_1, Z, \gamma_0))\Big) = 0$ for all B. By a conditioning argument similar to that used above, the projection is equal to the second term in the influence function

$$\frac{RZ(Y_2 - \mu_2^{(1)})}{\delta\pi^{(1)}(X_0, Y_1, X_1)} - d^T(X_0, Y_1, X_1)A_{(1)}^{-1}b_{(1)}\frac{\{R - \pi^{(1)}(X_0, Y_1, X_1)\}Z}{\delta}, \qquad (2.2.14)$$

which is (2.2.13) in this special case.

The preceding development assumes $\pi^{(1)}$ is correctly specified. If the postulated model is incorrect, then it is readily apparent that estimators derived from special

cases of (2.2.13), such as (2.2.7) and (2.2.14), may be inconsistent, as the influence function no longer has mean zero. However, in general, it may be shown that the "augmentation" in (2.2.8) induces the interesting property that, estimators derived from (2.2.8) will be consistent if either (i) the optimal choice of $g(O^*)$ is used, but $\pi(O^*)$ is misspecified or (ii) π is correctly specified, but the choice $g(O^*)$ does not correspond to the optimal one but $\pi(O^*)$ is correctly specified. This follows because under either (i) or (ii), the resulting influence function still turns out to have mean zero. Such an estimator will of necessity be inefficient, as its influence function is no longer the optimal one. This property has been referred to as "double robustness" (e.g., Scharfstein, Robins, and Rotnitzky 1999, sec.; van der Laan and Robins 2003 sec. 1.6). In the case of the pretest-posttest model, this corresponds in (2.2.12), for example, to (i) taking $h^{(1)}(X_0, Y_1) = E(Y_2|X_0, Y_1, Z = 1) - \mu_2^{(1)}$ and $q^{(1)}(X_0, Y_1, X_1) = E(Y_2|X_0, Y_1, X_1, Z = 1) - \mu_2^{(1)}$ but specifying $\pi^{(1)}(X_0, Y_1, X_1)$ incorrectly or (ii) taking $h^{(1)}$ and $q^{(1)}$ to be something other than these choices but positing a model for $\pi^{(1)}(X_0, Y_1, X_1)$ corresponding to the truth. Of course, in practice, one would not know these conditional expectations, so they would presumably have to be estimated somehow. This is discussed in the next chapter.

The same considerations outlined here lead to influence functions for estimators for $\mu_2^{(0)}$ of forms similar to those for $\mu_2^{(1)}$ in particular, the efficient influence function is of the form (2.2.12) with Z, $\pi^{(1)}$, $h^{(1)}$, $q^{(1)}$, and δ in the denominators replaced by $1 - Z$, $\pi^{(0)}$, $h^{(0)} = E(Y_2|X_0, Y_1, Z = 0) - \mu^{(0)}$, $q^{(0)} = E(Y_2|X_0, Y_1, X_1, Z = 0) - \mu_2^{(0)}$,

and $(1-\delta)$, respectively, with similar modifications in the case of (2.2.13):

$$\frac{RZ(Y_2-\mu_2^{(0)})}{(1-\delta)\pi(Y_1,X_0,X_1,Z=0)}+\frac{(Z-\delta)}{1-\delta}\{E(Y_2|Y_1,X_0,Z=0)-\mu_2^{(0)}\}$$
$$-\quad\frac{R-\pi(Y_1,X_0,X_1,Z=0)}{\delta\pi(Y_1,X_0,X_1,Z=0)}E(Y_2|Y_1,X_0,X_1,Z=0)-\mu_2^{(0)};$$

The optimal influence function for the treatment effect β is the difference of the optimal influence function for the treatment mean and for the control mean, which is equal to

$$\frac{RZ(Y_2-\mu_2-\beta)}{\delta\pi(Y_1,X_0,X_1,Z=1)}-\frac{R(1-Z)(Y_2-\mu_2)}{(1-\delta)\pi(Y_1,X_0,X_1,Z=0)}$$
$$-\quad(Z-\delta)\left\{\frac{(E(Y_2|Y_1,X_0,Z=1)-\mu_2^{(1)})}{\delta}+\frac{(E(Y_2|Y_1,X_0,Z=0)-\mu_2^{(0)})}{1-\delta}\right\}$$
$$-\quad\frac{R-\pi(Y_1,X_0,X_1,Z=1)}{\delta\pi(Y_1,X_0,X_1,Z=1)}(E(Y_2|Y_1,X_0,X_1,Z=1)-\mu_2^{(1)})$$
$$+\quad\frac{R-\pi(Y_1,X_0,X_1,Z=0)}{(1-\delta)\pi(Y_1,X_0,X_1,Z=0)}(E(Y_2|Y_1,X_0,X_1,Z=0)-\mu_2^{(0)}). \qquad (2.2.15)$$

Chapter 3

Practical Implementation

3.1 Full data

To exploit (2.1.6), one must identify estimators for β with this influence function, which will then be "optimal" as described in chapter 2. This may be accomplished by finding estimators for $\mu_2^{(1)} = \mu_2 + \beta$ and $\mu_2^{(0)} = \mu_2$ with the influence functions in (2.1.6) and taking their difference. An obvious complication is the involvement of the unknown conditional expectations $E(Y_2|Y_1, Z = c)$, $c = 0, 1$, which depend on the unspecified joint distribution of the observed data; thus, a way of deducing these quantities is required. One might use a form of nonparametric smoothing to estimate $E(Y_2|Y_1, Z = c)$ or fit specific parametric models based on inspection of plots like those in Figure 1.1 (Robins et al., 1994). Nonparametric estimators typically do not attain usual parametric $n^{-1/2}$-convergence rates, raising concern that effects of such smoothing will degrade performance of the estimator for β in small samples relative to

that achieved with a correct, parametric specification. For larger n, such smoothing may be a viable alternative; however, if baseline covariates X_0 are incorporated, multi-dimensional smoothing is required, which may be prohibitive if $\dim(X_0)$ is large. On the other hand, although the estimator for β will still be consistent and asymptotically normal as a member of the general class if the chosen parametric form is incorrect, the resulting estimator no longer need have the optimal influence function, so could in fact be inferior to the "popular" estimators. Choosing a parametric model can be tricky; for the ACTG 175 data, the nature of the "true" relationship is ambiguous, and Figure 1.1 suggests several plausible parametric models for each group, e.g., a linear, quadratic, or nonlinear (exponential) function. Regardless, one is still faced with the issue of deriving appropriate estimators for $\mu_2^{(1)}$ and $\mu_2^{(0)}$.

We propose a strategy that may be regarded as a "compromise" between fully nonparametric smoothing and parametric modeling and that leads straightforwardly to a general form of an estimator for β that will improve on the "popular" ones and has the optimal influence function under conditions we elucidate shortly. The approach is based on restricting the search for estimators for β to those with influence functions of the form

$$\frac{Z_i(Y_2 - \mu_2 - \beta)}{\delta} - \frac{(1 - Z_i)(Y_2 - \mu_2)}{1 - \delta} + (Z - \delta) f^T(Y_1)\alpha, \quad \alpha \in \Re^k, \tag{3.1.1}$$

where $f(Y_1) = \{f_1(Y_1), \ldots, f_k(Y_1)\}^T$ is a k-vector of basis functions. E.g., the $(k-1)$-order polynomial basis takes $f(Y_1) = \{1, Y_1, Y_1^2, \ldots, Y_1^{k-1}\}^T$; alternatively, one may choose a spline basis or discretization basis with $f(Y_1) = \{I(Y_1 < t_1), I(t_1 \leq Y_1 <$

$t_2), \ldots, I(t_{k-2} \leq Y_1 < t_{k-1}), I(Y_1 \geq t_{k-1})\}^T$ for $t_1 < t_2 < \cdots < t_{k-1}$. Thus, (3.1.1) may be viewed as restricting the search for h in (2.1.6) to the linear space spanned by $f(Y_1)$. If the basis is sufficiently rich so that this space is a good approximation to the space of all possible h, then the resulting estimators should be close to "optimal" if they are "optimal" within the restricted class.

We thus find the most efficient influence function in class (3.1.1), which follows by identifying α minimizing the variance of (3.1.1). Under our assumptions, the first two terms of (3.1.1) are uncorrelated, so this is equivalent to minimizing $\text{var}(A - B^T\alpha)$, where A corresponds to the first two terms and $B = -(Z - \delta)f(Y_1)$. This is an unweighted least squares problem, so that $\alpha^T = \text{cov}(A, B)\{\text{var}(B^T)\}^{-1}$, which may be shown to be $\alpha^T = -\{(1-\delta)\Sigma_{fY_2}^{(1)T} + \delta\Sigma_{fY_2}^{(0)T}\}\Sigma_{ff}^{-1}/\{\delta(1-\delta)\}$, where $\Sigma_{fY_2}^{(0)} = E\{f(Y_1)(Y_2 - \mu_2 - \beta)|Z = 1\}$, $\Sigma_{fY_2}^{(1)} = E\{f(Y_1)(Y_2 - \mu_2)|Z = 0\}$, and $\Sigma_{ff} = E\{f(Y_1)f^T(Y_1)\}$. In fact, this α may be found by the sum of separate regressions of each term on B. Thus, the optimal influence function in the class (3.1.1) is

$$\left\{ \frac{Z_i(Y_2 - \mu_2 - \beta)}{\delta} - \frac{(Z - \delta)\Sigma_{fY_2}^{(1)T}\Sigma_{ff}^{-1}f(Y_1)}{\delta} \right\} - \left\{ \frac{(1 - Z)(Y_2 - \mu_2)}{1 - \delta} + \frac{(Z - \delta)\Sigma_{fY_2}^{(0)T}\Sigma_{ff}^{-1}f(Y_1)}{1 - \delta} \right\} \tag{3.1.2}$$

$$= \frac{Z_i(Y_2 - \mu_2 - \beta)}{\delta} - \frac{(1 - Z)(Y_2 - \mu_2)}{1 - \delta} - (Z - \delta)\left\{ \frac{(1 - \delta)\Sigma_{fY_2}^{(1)} + \delta\Sigma_{fY_2}^{(0)}}{\delta(1 - \delta)} \right\}^T \Sigma_{ff}^{-1}f(Y_1) \tag{3.1.3}$$

Derivation of an estimator for β with influence function (3.1.3) is straightforward by considering the equivalent representation (3.1.2) to deduce estimators for each of $\mu_2^{(1)} = \mu_2 + \beta$ and $\mu_2^{(0)} = \mu_2$ and taking their difference. An estimator for $\mu_2^{(1)}$ with influence function equal to the first term in braces in (3.1.2) may be found by

equating the sample average of such terms to zero to obtain $\mu_2^{(1)} = \Big[n^{-1}\sum_{i=1}^n\{Z_iY_{2i} - (Z_i-\delta)\Sigma_{fY_2}^{(1)T}\Sigma_{ff}^{-1}f(Y_1)\}\Big]/\{n^{-1}\sum_{i=1}^n Z_i\}$, which suggests the estimator for $\mu_2^{(1)}$ found by substituting sample moment analogs for each quantity in this expression. Using similar calculations for the second term in braces in (3.1.2) to isolate $\mu_2^{(0)}$ and defining $S_{fY_2}^{(c)} = \sum_{i=1}^n I(Z_i = c)f(Y_{1i})(Y_{2i} - \overline{Y}_2^{(c)})$, $c = 0, 1$; $S_{ff} = \sum_{i=1}^n f(Y_{1i})f^T(Y_{1i})$; and $S_{fZ} = \sum_{i=1}^n (Z_i - n_1/n)f(Y_{1i})$, by taking the difference, we obtain the estimator for β

$$\widehat{\beta} = \overline{Y}_2^{(1)} - \overline{Y}_2^{(0)} - n\left(\frac{S_{fY_2}^{(1)}}{n_1^2} + \frac{S_{fY_2}^{(0)}}{n_0^2}\right)^T S_{ff}^{-1}S_{fZ}, \tag{3.1.4}$$

which may be shown to have influence function (3.1.3). It is straightforward to show that the variance of (3.1.3), and hence the large-sample variance of $n^{1/2}(\widehat{\beta} - \beta)$, is given by

$$\frac{\sigma_{22}^{(1)}}{\delta} + \frac{\sigma_{22}^{(0)}}{1-\delta} - \delta(1-\delta)\left(\frac{\Sigma_{fY_2}^{(1)}}{\delta} + \frac{\Sigma_{fY_2}^{(0)}}{1-\delta}\right)^T \Sigma_{ff}^{-1}\left(\frac{\Sigma_{fY_2}^{(1)}}{\delta} + \frac{\Sigma_{fY_2}^{(0)}}{1-\delta}\right), \tag{3.1.5}$$

which suggests the estimator for sampling variance of $\widehat{\beta}$ given by

$$\frac{S_{22}^{(1)}}{n_1^2} + \frac{S_{22}^{(0)}}{n_0^2} - n_1n_0\left(\frac{S_{fY_2}^{(1)}}{n_1^2} + \frac{S_{fY_2}^{(0)}}{n_0^2}\right)^T S_{ff}^{-1}\left(\frac{S_{fY_2}^{(1)}}{n_1^2} + \frac{S_{fY_2}^{(0)}}{n_0^2}\right), \tag{3.1.6}$$

where $S_{22}^{(c)} = \sum_{i=1}^n I(Z_i = c)(Y_{2i} - \overline{Y}_2^{(c)})^2$, $c = 0, 1$.

A modification is to use different sets of basis functions, $f^{(c)}(Y_1)$, $c = 0, 1$, say, for each term in (3.1.2). Alternatively, these developments suggest applying a similar approach to (2.1.6), representing $E(Y_2|Y_1, Z = c)$, $c = 0, 1$, by linear combinations of basis functions. It is straightforward to show that this leads to the same class of estimators represented by (3.1.1), suggesting that, in practice, insight may be gained into the choice of basis by examining plots such as Figure 1.1. Thus, the approach may be

viewed as intermediate to completely nonparametric and fully parametric estimation of the $E(Y_2|Y_1, Z = c)$, approximating these quantities by a finite-dimensional, flexible form emphasizing the predominant trend apparent in the data. From the derivation of the optimal α, to achieve efficiency gains, it is essential that this be carried out by unweighted regression, even if $\text{var}(Y_2|Y_1, Z = c)$, $c = 0, 1$, is not constant with respect to Y_1. When the true conditional expectations follow exactly such a form and k and the basis functions are correctly chosen, the method will yield asymptotically the most efficient estimator for β; otherwise, we expect gains over the "popular" estimators as long as the basis approximates a broad range of relationships. In the next chapter, we demonstrate that close-to-"optimal" inference is obtained when the true relationship is nonlinear but the basis representation captures its salient features.

The influence function and its variance depend on moments of squares and crossproducts of elements of $f(Y_1)$ through Σ_{ff} and the covariances of $Y_2^{(c)}$, $c = 0, 1$, with elements of $f(Y_1)$ through $\Sigma_{fY_2}^{(c)}$, which are estimated in (3.1.4) and (3.1.6) by sample analogs. Thus, e.g., the quadratic basis $f(Y_1) = (1, Y_1, Y_1^2)^T$ used in the next chapter involves estimation of not only σ_{11}, $\sigma_{12}^{(c)}$, and $\sigma_{22}^{(c)}$ but also of $\text{cov}(Y_1^2, Y_2^{(c)})$, $c = 0, 1$, and the coefficients of skewness and excess kurtosis of the distribution of Y_1. This suggests that attempting to gain efficiency over "popular" estimators in this way with small sample sizes is unwise. However, in situations such as large clinical trials, this approach may be fruitful. The simulation evidence in Chapter 4 indicates that impressive efficiency gains are possible in the moderate-to-large sample sizes where the improved estimators are expected to perform well.

When baseline covariates are available, the strategy is immediately extended by replacing $f(Y_1)$ in (3.1.1) by a k-vector of basis functions $f(X_0, Y_1)$; e.g., one might choose a polynomial basis including interactions of powers of Y_1 with elements of X_0.

3.2 Missing at random follow-up.

The form of the efficient influence function given in 2.2.15is a natural starting point from which to derive estimators with good properties. In addition to postulating a parametric model for π, the analyst must characterize somehow the conditional quantities $E(Y_2|X_0, Y_1, Z = c)$ and $E(Y_2|X_0, Y_1, X_1, Z = c)$, $c = 0, 1$. This may be challenging, as we now demonstrate.

One possibility is to adopt parametric models for the conditional expectations based on usual regression considerations, fit these, obtain predicted values for each subject, substitute in (2.2.11), and deduce estimators as described below from the resulting expressions for each i. Because of the assumption of MAR, $E(Y_2|X_0, Y_1, X_1, Z, R)$ does not depend on R; thus, $E(Y_2|X_0, Y_1, X_1, Z) = E(Y_2|X_0, Y_1, X_1, Z, R = 1)$, implying that such a model may be postulated and fitted based only on the complete cases. Let $\widehat{e}_{q(c)i}$ be the resulting estimator for $E(Y_{2i}|X_{0i}, Y_{1i}, X_{1i}, Z_i = c)$, $c = 0, 1$, for subject i. Considerations for $E(Y_2|X_0, Y_1, Z)$ are trickier. Ideally, the chosen model for this quantity must be compatible with that for $E(Y_2|X_0, Y_1, X_1, Z)$, as $E(Y_2|X_0, Y_1, Z) = E\{E(Y_2|X_0, Y_1, X_1, Z)|X_0, Y_1, Z\}$. Several practical strategies are possible, although none is guaranteed to achieve this property and hence yield an

efficient estimator for $\mu_2^{(1)}$. One approach is to independently adopt a model for $E(Y_2|X_0, Y_1, Z)$ directly and hope that it is "close enough" to be "approximately compatible." E.g., if $E(Y_2|X_0, Y_1, X_1, Z)$ is linear in all its arguments, one may be comfortable choosing a model for $E(Y_2|X_0, Y_1, Z)$ that is also linear. If all of X_0, Y_1, X_1 are continuous, an assumption of joint normality may be a reasonable approximation, in which case standard results may be used to deduce both models; as these variables are likely to be a mix of continuous and discrete components, this strategy may be of limited practical utility. Note that for any chosen model for $E(Y_2|X_0, Y_1, Z)$, it is no longer appropriate to fit the model based on the complete cases only. Thus, fitting would need to be carried out by a procedure that takes account of the fact that Y_2 is MAR; e.g., an IWCC version of standard regression techniques. Following any such approach, predicted values $\widehat{e}_{h(c)i}$, say, would follow from the resulting estimator for $E(Y_{2i}|X_{0i}, Y_{1i}, , Z_i = c)$, $c = 0, 1$.

Alternatively, one might use the relationship $E(Y_2|X_0, Y_1, Z) = E\{E(Y_2|X_0, Y_1, X_1, Z)|X_0, Y_1, Z\}$. For example, if X_1 is one-dimensional, a distributional model for $X_1|X_0, Y_1, Z$ might be postulated and fitted based on the $(X_{0i}, Y_{1i}, X_{1i}, Z_i)$, $i = 1, \ldots, n$, which are observed for all subjects; integration with respect to this model would lead to the desired conditional quantities for $c = 0, 1$. When X_1 is binary, as in the simulations in section 4.2, a logistic model for $P(X_1 = 1|X_0, Y_1, Z)$ may be used, and such integration is straightforward. Instead, one might invoke an empirical approximation; for example, one might obtain the predicted value $\widehat{e}_{h(c)i}$, $c = 0, 1$ for each i by averaging estimates of $E(Y_{2i}|X_{0i}, Y_{1i}, X_{1j}, Z_i = c)$ over subjects j sharing

the same values for (X_0, Y_1, Z) as i; this would likely be feasible only in specialized circumstances. A cruder version of this strategy would be to average over all X_{1j} for j in the same group as i; of course, this would yield the desired result only if X_1 is conditionally independent of (X_0, Y_1) given Z.

However predicted values $\widehat{e}_{q(c)i}$, $\widehat{e}_{h(c)i}$, $c = 0, 1$ are deduced for each i, an estimator for $\mu_2^{(1)}$ may be constructed by setting the sum for $i = 1, \ldots, n$ of terms of the form in (2.2.12) for each subject to zero; substituting $\widehat{e}_{h(1)i} - \mu_2^{(1)}$ and $\widehat{e}_{q(1)i} - \mu_2^{(1)}$ for $h^{(1)}(X_{0i}, Y_{1i})$ and $q^{(1)}(X_{0i}, Y_{1i}, X_{1i})$, respectively; and solving for $\mu_2^{(1)}$; an analogous development is possible for $\mu_2^{(0)}$. Letting $\widehat{\pi}_i^{(c)} = \pi^{(c)}(X_{0i}, Y_{1i}, X_{1i}; \widehat{\gamma})$ and $\widehat{n}_{RZ(c)} = \sum_{i=1}^n R_i I(Z_i = c)/\widehat{\pi}_i^{(c)}$ and substituting $\widehat{\delta} = n_1/n$ for δ, simple algebra yields the estimator for β given by $\widehat{\mu}_2^{(1)} - \widehat{\mu}_2^{(0)}$, where $\widehat{\mu}_2^{(1)} = n_1^{-1}\{\sum_{i=1}^n R_i Z_i Y_{2i}/\widehat{\pi}_i^{(1)} - \sum_{i=1}^n (Z_i - \widehat{\delta})\widehat{e}_{h(1)i} - \sum_{i=1}^n (R_i - \widehat{\pi}_i^{(1)}) Z_i \widehat{e}_{q(1)i}/\widehat{\pi}_i^{(1)}\}$ and $\widehat{\mu}_2^{(0)} = n_0^{-1}\{\sum_{i=1}^n R_i(1 - Z_i) Y_{2i}/\widehat{\pi}_i^{(0)} - \sum_{i=1}^n (Z_i - \widehat{\delta})\widehat{e}_{h(0)i} - \sum_{i=1}^n (R_i - \widehat{\pi}_i^{(1)})(1 - Z_i)\widehat{e}_{q(1)i}/\widehat{\pi}_i^{(0)}\}$.

The asymptotic variance for $\widehat{\beta} = \widehat{\mu}_2^{(1)} - \widehat{\mu}_2^{(0)}$ can be obtained from the expectation of the square of 2.2.15, which, after some simplification yields,

$$
\begin{aligned}
& E\left\{ \frac{(Y_2 - \mu_2^{(1)})^2}{\pi^{(1)}(X_0, Y_1, X_1)\delta} \middle| Z = 1 \right\} + E\left\{ \frac{(Y_2 - \mu_2^{(0)})^2}{\pi^{(0)}(X_0, Y_1, X_1)(1 - \delta)} \middle| Z = 0 \right\} \\
& - \quad \delta(1 - \delta) E\left\{ \frac{E(Y_2 | X_0, Y_1, Z = 1) - \mu_2^{(1)}}{\delta} + \frac{E(Y_2 | X_0, Y_1, Z = 0) - \mu_2^{(0)}}{1 - \delta} \right\}^2 \\
& - \quad \sum_{c=0,1} \left(\frac{I(c = 1)}{\delta} + \frac{I(c = 0)}{1 - \delta} \right) E\left[\frac{1 - \pi^{(c)}}{\pi^{(c)}} \{ E(Y_2 | X_0, Y_1, X_1, Z = c) - \mu_2^{(c)} \} \right].
\end{aligned}
\tag{3.2.7}
$$

If the models used to obtain the predicted values and those for $\pi^{(c)}$, $c = 0, 1$,

are correctly specified, then the resulting estimator should be efficient in the sense described earlier. Although additional regression parameters must be estimated, because of the geometry, there is no effect asymptotically. The "double robustness" property discussed at the end of Chapter 2.2 ensures that consistent estimators for β and $\mu_2^{(c)}$, $c = 0, 1$, will be obtained as long as either set of models is correct; however, efficiency is no longer guaranteed. To address possible misspecification of models for $E(Y_2|X_0, Y_1, X_0, Z)$ and $E(Y_2|X_0, Y_1, Z)$, one might contemplate using a form of nonparametric smoothing to estimate these quantities; e.g., locally weighted polynomial smoothing (Cleveland, Gross, and Shyu 1993) or generalized additive modeling (Hastie and Tibshirani 1990). However, this involves several difficulties. If X_0, X_1 are high-dimensional, such smoothing is likely to be problematic. Even if not, usual fitting a nonparametric model for $E(Y_2|X_0, Y_1, Z)$ would need to be modified to take into account that Y_2 is MAR, and one would still face the issue of compatibility. If instead an estimate of $E(Y_2|X_0, Y_1, Z)$ were derived from the nonparametric fit of $E(Y_2|X_0, Y_1, X_1, Z)$, integration following the smoothing would be required.

To summarize, despite the relative simplicity of the pretest-posttest model, except in low-dimensional, simple situations, identifying estimators with the efficient influence function may be a daunting task in practice. In the next section, we consider an alternative approach.

3.2.1 Estimators Based on a Restricted Class of Influence Functions

A strategy for finding estimators that circumvents the above difficulties and that has been used successfully in other contexts (e.g., Bang and Tsiatis 2000; Leon et al. 2003) is to restrict attention to a judiciously-chosen subclass of influence functions of the form (2.2.13) and find the efficient estimator within this subclass. In particular, we consider the subclass of (2.2.13) with $h^{(1)}(X_0, Y_1) = \alpha^T f(X_0, Y_1)$ and $q^{(1)}(X_0, Y_1, X_1) = \eta^T \ell(X_0, Y_1, X_1)$, where $\alpha \in \Re^k$; $\eta \in \Re^m$; and $f(Y_1, X_0) = \{f_1(X_0, Y_1), \ldots, f_k(X_0, Y_1)\}^T$ and $\ell(X_0, Y_1, X_1) = \{\ell_1(X_0, Y_1, X_1), \ldots, \ell_m(X_0, Y_1, X_1)\}^T$ are k- and m-vectors of basis functions, respectively, chosen by the analyst. E.g., the polynomial basis $f(X_0, Y_1)$ of order 2 is $(1, X_0, Y_1, X_0Y_1, X_0^2, Y_1^2)^T$, and $k = 6$, and similarly for $\ell(X_0, Y_1, X_1)$. Other choices of bases are discussed by Leon et al. (2003). Thus, the subclass corresponds to restricting the choice $h^{(1)}$ and $q^{(1)}$ in (2.2.13) to the linear spaces spanned by the selected bases $f(X_0, Y_1)$ and $\ell(X_0, Y_1, X_1)$. The rationale is that, if the bases are sufficiently flexible to provide a good approximation to the spaces of all possible $h^{(1)}$ and $q^{(1)}$, then estimators derived from the resulting influence function should be "close" to "optimal" if they are efficient within the subclass.

Accordingly, we propose to find the most efficient influence function in the re-

stricted class

$$\left[\frac{RZ(Y_2-\mu_2-\beta)}{\delta\pi^{(1)}(X_0,Y_1,X_1)}-b_{(1)}^TA_{(1)}^{-1}d(Y_1,X_1,X_0)\frac{\{R-\pi^{(1)}(X_0,Y_1,X_1)\}Z}{\delta}\right]-$$
$$\frac{(Z-\delta)}{\delta}f^T(X_0,Y_1)\alpha-\frac{\{R-\pi^{(1)}(X_0,Y_1,X_1)\}Z}{\delta}\left\{\frac{\ell^T(Y_1,X_1,X_0)}{\pi^{(1)}(X_0,Y_1,X_1)}-\right.$$
$$\left.d^T(X_0,Y_1,X_1)A_{(1)}^{-1}b_{\ell(1)}\right\}\eta, \tag{3.2.8}$$

where $b_{\ell(1)}=E[\{1-\pi^{(1)}(X_0,Y_1,X_1)\}d(X_0,Y_1,X_1)\ell^T(Y_1,X_1,X_0)|Z=1]$. To find the efficient influence function in class (3.2.8), we must identify α and η such that the variance of (3.2.8) is minimized. (3.2.8) is of form $A-B_1^T\alpha-B_2^T\eta$, and it may be shown that B_1 and B_2 are uncorrelated. Thus, the values $\alpha^{(1)}$ and $\eta^{(1)}$, say, minimizing this variance may be found from separate unweighted regressions of A on B_1 and B_2 and are given by $\alpha^{(1)}=\Sigma_{ff}^{-1}\Sigma_{fY_2}^{(1)}$, and $\eta^{(1)}=(\Sigma_{\ell\ell}^{(1)}-b_{\ell(1)}^TA_{(1)}^{-1}b_{\ell(1)})^{-1}(\Sigma_{\ell Y_2}^{(1)}-b_{\ell(1)}^TA_{(1)}^{-1}b_{(1)})$, where $\Sigma_{ff}=E\{f(Y_1)f^T(Y_1)\}$, $\Sigma_{fY_2}^{(1)}=E\{f(Y_1)(Y_2-\mu_2^{(1)})|Z=1\}$, $\Sigma_{\ell\ell}^{(1)}=E[\{1-\pi^{(1)}(X_0,Y_1,X_1)\}\ell(X_0,Y_1,X_1)\ell^T(X_0,Y_1,X_1)/\pi^{(1)}(X_0,Y_1,X_1)|Z=1]$, and $\Sigma_{\ell Y_2}^{(1)}=E[(Y_2-\mu_2^{(1)})\{1-\pi^{(1)}(X_0,Y_1,X_1)\}\ell(X_0,Y_1,X_1)/\pi^{(1)}(X_0,Y_1,X_1)|Z=1]$.

We thus desire to deduce estimators for $\mu_2^{(1)}$ with influence function (3.2.8), with $\alpha=\alpha^{(1)}$ and $\eta=\eta^{(1)}$. Recall from the discussion following (2.2.13) that, to find estimators with influence functions in this class, it suffices to find estimators with influence functions of the form (2.2.12) and substitute the ML estimator for γ. Accordingly, we consider influence functions of form (2.2.12) with $h^{(1)}(X_0,Y_1)$ and $q^{(1)}(X_0,Y_1,X_1)$ replaced by $f^T(X_0,Y_1)\alpha^{(1)}$ and $\ell^T(X_0,Y_1,X_1)\eta^{(1)}$, respectively. An estimator may be found by setting the sum of terms of the form of (2.2.12) with these choices of $h^{(1)}$ and $q^{(1)}$ for $i=1,\ldots,n$ equal to zero, solving for $\mu_2^{(1)}$, and

substituting estimators for quantities that appear in the resulting expressions. Write $\widehat{\pi}_i^{(1)} = \pi^{(1)}(X_{0i}, Y_{1i}, X_{1i}; \widehat{\gamma})$ and $\widehat{n}_{RZ(1)} = \sum_{i=1}^n R_i Z_i / \widehat{\pi}_i^{(1)}$ as before, $f_i = f(X_{0i}, Y_{1i})$, $\ell_i = \ell(X_{0i}, Y_{1i}, X_{1i})$, and $d_i = d(X_{0i}, Y_{1i}, X_{1i})$. Now define $\widehat{\Sigma}_{ff} = n^{-1}\sum_{i=1}^n f_i f_i^T$, $\widehat{b}_{\ell(1)} = n_1^{-1}\sum_{i=1}^n Z_i(1-\widehat{\pi}_i^{(1)}) d_i \ell_i^T$, $\widehat{\Sigma}_{\ell\ell}^{(1)} = n_1^{-1}\sum_{i=1}^n Z_i(1-\widehat{\pi}_i^{(1)})\ell_i\ell_i^T/\widehat{\pi}_i^{(1)}$, $\widehat{A}^{(1)} = n_1^{-1}\sum_{i=1}^n Z_i(1-\widehat{\pi}_i^{(1)})\widehat{\pi}_i^{(1)} d_i d_i^{-1}$, $\widehat{c}_{f(1)} = n_1^{-1}\sum_{i=1}^n Z_i f_i$, $\widehat{c}_{\ell(1)} = n_1^{-1}\sum_{i=1}^n Z_i(1-\widehat{\pi}_i^{(1)})\ell_i/\widehat{\pi}_i^{(1)}$, $\widehat{c}_{d(1)} = n_1^{-1}\sum_{i=1}^n Z_i(1-\widehat{\pi}_i^{(1)}) d_i$, $\widehat{c}_{fY_2(1)} = \widehat{n}_{RZ(1)}^{-1}\sum_{i=1}^n R_i Z_i Y_{2i} f_i/\widehat{\pi}_i^{(1)}$, $\widehat{c}_{\ell Y_2(1)} = \widehat{n}_{RZ(1)}^{-1}$ $\sum_{i=1}^n R_i Z_i Y_{2i}\ell_i(1-\widehat{\pi}_i^{(1)})/\widehat{\pi}_i^{(1)\,2}$, $\widehat{c}_{dY_2(1)} = \widehat{n}_{RZ(1)}^{-1}\sum_{i=1}^n R_i Z_i Y_{2i}(1-\widehat{\pi}_i^{(1)}) d_i/\widehat{\pi}_i^{(1)}$, and $\widehat{\delta} = n_1/n$. Then, defining $\widehat{\Sigma}_{\ell\ell\bullet bA(1)} = \widehat{\Sigma}_{\ell\ell}^{(1)} - \widehat{b}_{\ell(1)}^T \widehat{A}_{(1)}^{-1}\widehat{b}_{\ell(1)}$, the resulting estimator $\widehat{\mu}_2^{(1)}$ is

$$\frac{\displaystyle\sum_{i=1}^n R_i Z_i Y_{2i}/\widehat{\pi}_i^{(1)} - \sum_{i=1}^n (Z_i - \widehat{\delta}) f_i^T \widehat{\Sigma}_{ff}^{-1}\widehat{c}_{fY_2(1)} - \left\{\sum_{i=1}^n (R_i - \widehat{\pi}_i^{(1)}) Z_i \ell_i^T/\widehat{\pi}_i^{(1)}\right\}\widehat{\Sigma}_{\ell\ell\bullet bA(1)}^{-1}}{\widehat{n}_{RZ(1)} - \displaystyle\sum_{i=1}^n (Z_i - \widehat{\delta}) f_i^T \widehat{\Sigma}_{ff}^{-1}\widehat{c}_{f(1)} - \left\{\sum_{i=1}^n (R_i - \widehat{\pi}_i^{(1)}) Z_i \ell_i^T/\widehat{\pi}_i^{(1)}\right\}\widehat{\Sigma}_{\ell\ell\bullet bA(1)}^{-1}} \frac{(\widehat{c}_{\ell Y_2(1)} - \widehat{b}_{\ell(1)}^T \widehat{A}_{(1)}^{-1}\widehat{c}_{dY_2(1)})}{(\widehat{c}_{\ell(1)} - \widehat{b}_{\ell(1)}^T \widehat{A}_{(1)}^{-1}\widehat{c}_{d(1)})} \tag{3.2.9}$$

Note that the quotient of the leading terms in the numerator and denominator of (3.2.9) is the IWCC estimator for $\mu_2^{(1)}$, where γ is estimated, so that (3.2.9) may be interpreted as a modification of this simple approach to increase efficiency.

An entirely similar development is possible for $\mu_2^{(0)}$. Define $\Sigma_{fY_2}^{(0)}$, $\Sigma_{\ell Y_2}^{(0)}$, $\Sigma_{\ell\ell}^{(0)}$, $b_{\ell(0)}$, $b_{(0)}$, and $A_{(0)}$ analogous to the preceding expressions, where $\mu_2^{(1)}$ and $\pi^{(1)}$ are replaced by $\mu_2^{(0)}$ and $\pi^{(0)}$, respectively; conditioning is with respect to $Z = 0$; and the vector $d(X_0, Y_1, X_1)$ may be different depending on the model $\pi^{(0)}$. One may also define analogously the sample quantities $\widehat{b}_{\ell(1)}$, $\widehat{\Sigma}_{\ell\ell}^{(0)}$, $\widehat{A}_{(0)}$, $\widehat{c}_{f(0)}$, $\widehat{c}_{\ell(0)}$, $\widehat{c}_{d(0)}$, $\widehat{c}_{fY_2(0)}$, $\widehat{c}_{\ell Y_2(0)}$, and $\widehat{c}_{dY_2(0)}$, where Z_i is replaced by $(1 - Z_i)$ and $\widehat{\pi}_i^{(1)}$ is replaced by $\widehat{\pi}_i^{(0)} = \widehat{\pi}^{(0)}(X_{0i}, Y_{1i}, X_{1i}; \widehat{\gamma})$, and γ in $\pi^{(0)}$ is estimated based on the data from control sub-

jects only. The estimator $\widehat{\mu}_2^{(0)}$ is defined as in (3.2.9), substituting these expressions in the obvious manner and replacing Z_i, $\widehat{\delta}$, and $\widehat{\pi}_i^{(1)}$ by $(1 - Z_i)$, $(1 - \widehat{\delta})$, and $\widehat{\pi}_i^{(0)}$, respectively, in the remaining quantities.

The estimator for β is thus given by $\widehat{\beta} = \widehat{\mu}_2^{(1)} - \widehat{\mu}_2^{(0)}$, and has influence function equal to the difference of the individual influence functions, which have the form (3.2.8) with the optimal choices for α and η, with the appropriate substitutions for the control mean. The asymptotic variance of $\widehat{\beta}$ is the variance of this difference, which may be shown to be

$$
\begin{aligned}
& E\left\{ \frac{(Y_2 - \mu_2^{(1)})^2}{\pi^{(1)}(X_0, Y_1, X_1)\delta} \middle| Z = 1 \right\} + E\left\{ \frac{(Y_2 - \mu_2^{(0)})^2}{\pi^{(0)}(X_0, Y_1, X_1)(1-\delta)} \middle| Z = 0 \right\} \\
& - \quad \delta(1-\delta)\left(\frac{\Sigma_{fY_2}^{(1)}}{\delta} + \frac{\Sigma_{fY_2}^{(0)}}{1-\delta} \right)^T \Sigma_{ff}^{-1} \left(\frac{\Sigma_{fY_2}^{(1)}}{\delta} + \frac{\Sigma_{fY_2}^{(0)}}{1-\delta} \right) \\
& - \quad \sum_{c=0,1} \left(\frac{I(c=1)}{\delta} + \frac{I(c=0)}{1-\delta} \right) (\Sigma_{(\ell Y_2)}^{(c)} - b_{\ell(c)}^T A_{(c)}^{-1} b_{(c)})^T (\Sigma_{\ell\ell}^{(c)} - b_{\ell(c)}^T A_{(c)}^{-1} b_{\ell(c)})^{-1} \\
& \qquad (\Sigma_{(\ell Y_2)}^{(c)} - b_{\ell(c)}^T A_{(c)}^{-1} b_{(c)}) - \frac{1}{\delta} b_{(1)}^T A_{(1)}^{-1} b_{(1)} - \frac{1}{1-\delta} b_{(0)}^T A_{(0)}^{-1} b_{(0)}. \qquad (3.2.10)
\end{aligned}
$$

This may be estimated by replacing all quantities in each term after the first two by the estimators defined above, and estimating the first two terms by $(\widehat{\delta}\widehat{n}_{RZ(1)})^{-1} \sum_{i=1}^n R_i Z_i (Y_{2i} - \widehat{\mu}_2^{(1)})^2 / \widehat{\pi}_i^{(1)\,2}$ and $\{(1-\widehat{\delta})\widehat{n}_{RZ(0)}\}^{-1} \sum_{i=1}^n R_i (1 - Z_i)(Y_{2i} - \widehat{\mu}_2^{(0)})^2 / \widehat{\pi}_i^{(0)\,2}$, respectively.

One may use the same basis functions for each intervention group, as shown above, or choose different bases $f^{(c)}(X_0, Y_1)$ and $\ell^{(c)}(X_0, Y_1, X_1)$ for $c = 0, 1$. One may view the approach as an attempt to approximate the quantities $E(Y_2 | X_0, Y_1, Z = c)$ and $E(Y_2 | X_0, Y_1, X_1, Z = c)$ by flexible, finite-dimensional representations that attempt to capture the predominant relationships among variables. When the true conditional

expectations follow exactly the form indicated by the bases, with k and m correctly chosen, the asymptotically most efficient estimator will be obtained. Otherwise, if the bases approximate a range of potential relationships, we expect efficiency gains over methods such as simple IWCC. Evidence supporting this contention is presented in the next chapter.

The form of (3.2.9), requiring estimation of complicated quantities, some by IWCC estimation, suggests that it may be unwise to adopt this approach in small samples, as there is a potential that these quantities may be poorly estimated, which could lead to degradation of performance. However, in large studies, the proposed estimator offers a feasible approach to consistent inference when Y_2 is MAR that should offer efficiency gains over simpler methods.

Chapter 4

Simulation evidence

4.1 Full data.

We carried out several simulation studies, and report here on results for four scenarios, each involving 5000 Monte Carlo replications, $\beta = 0.5$, $\delta = 0.5$, and $Y_1 \sim \mathcal{N}(\mu_1 = 0, \sigma_{11} = 1)$. We estimated β by the ANCOVA I and II; two-sample t-test; and paired t-test estimators; (3.1.4) with quadratic polynomial basis, denoted QUAD; and the estimator formed by estimating $E(Y_2|Y_1, Z = c)$, $c = 0, 1$, via locally weighted polynomial smoothing using `proc loess` in SAS (SAS Institute, 2000) using quadratic polynomials, substituting in (2.1.6), finding estimators for $\mu_2^{(0)}$ and $\mu_2^{(1)}$ by equating sample averages of each term in (2.1.6) to zero, and taking their difference, denoted LOESS. For "popular" estimators, standard errors were obtained both by substituting sample moments in the asymptotic variance formulæ suggested by their influence functions, given explicitly in Section 2 of Yang and Tsiatis (2001), and us-

ing the "usual" expressions for these estimators, e.g., for ANCOVA I and II obtained from standard OLS formulæ. For QUAD and LOESS, standard errors were obtained from (3.1.6) and the asymptotic formula. Nominal 95% Wald confidence intervals for β were constructed as the estimate ± 1.96 times the asymptotic-formula standard error.

Follow-up responses for the first two scenarios were generated from the quadratic model

$$Y_{2i} = (\mu_2 + \beta Z_i) + \beta_1(Y_{1i} - \mu_1) + \beta_2\{(Y_{1i} - \mu_1)^2 - \sigma_{11}\} + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}((0,1) \quad (4.1.1)$$

with $\mu_2 = -0.25$. Normality of baseline and follow-up may be considered the "most favorable" distribution for the "popular" estimators, which are often thought to be predicated on normality, although from Section 2.1 these estimators are consistent more generally. Situation Q1 is based on (4.1.1) with $(\beta_1, \beta_2) = (0.5, 0.4)$, yielding a discernible curvilinear relationship between baseline and follow-up, depicted in Figure 4.1(a). The "popular" estimators assume a linear relationship, thus, the linear correlation of 0.40 between baseline and follow-up in each group is of interest. The second situation, Q2, exemplified in Figure 4.1(b), with $(\beta_1, \beta_2) = (0.1, 0.1)$, involves "low" correlation of 0.10 in each group. Tables 4.1 and 4.2 shows results for $n = 100$, 500, which for $n = 100$ are similar to those of Yang and Tsiatis (2001), who used (4.1.1) but with interactions between baseline and treatment and who mislabelled Monte Carlo standard deviation and standard error estimates as "variance." In all cases, bias is negligible, standard error estimates are reliable, and coverage probabilities are close to the nominal level. ANCOVA I/II are equivalent; for Q2, the

two-sample t-test performs well, and the paired t-test is inefficient for both scenarios, as expected. Most striking is the considerable gain in efficiency attained by QUAD over the "popular" approaches in the realistic situation of Q1, with a moderate degree of curvilinear association. The LOESS method is virtually equivalent to QUAD for $n = 500$ in both Q1 and Q2 but for the smaller $n = 100$ exhibits some loss of efficiency, perhaps reflecting the concern raised in Section 3.1. Overall, the proposed approach, which seeks to enhance performance by exploiting the nature of the relationship between baseline and follow-up, can dramatically improve precision. From Table 4.1, as expected, little is to be gained when this relationship is weak (Q2).

In Q1 and Q2, the basis functions coincided with the true form of $E(Y_2|Y_1, Z)$. To investigate performance for a more complicated relationship, we generated data from

$$Y_{2i} = \beta_0 + \beta Z_i + e^{\beta_1 + \beta_2 Y_{1i}} + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0, 1), \tag{4.1.2}$$

where now μ_2 depends on $(\beta_0, \beta_1, \beta_2)$. In the first situation, N1, $(\beta_0, \beta_1, \beta_2) = (-4.0, 1.0, 0.5)$, resulting in curvature typified by Figure 4.1(c), with "high" linear correlation between baseline and follow-up of about 0.80 in each group. Situation N2, with $(\beta_0, \beta_1, \beta_2) = (-4.0, 1.4, 0.1)$, produces haphazard scatterplots as in Figure 4.1(d) and "weak" correlation of roughly 0.30. In (4.1.2), $E(Y_2|Y_1, Z)$ is not quadratic; thus, as an "ideal" benchmark, we also estimated β by taking the true $E(Y_2|Y_1, Z)$ to be known and finding the optimal estimator based on (2.1.6), denoted as BENCHMARK in Tables 4.3 and 4.4 . For N1, LOESS and QUAD perform well relative to the unachievable "ideal" and offer appreciable gains in efficiency relative

to the "popular" estimators, despite "high" correlation, although again LOESS is less precise when $n = 100$. Not unexpectedly, for N2, with no discernible relationship, the BENCHMARK, QUAD, and LOESS estimators offer no improvement over the best "linear" estimators.

It is natural to wonder if these efficiency gains translate to increased power for testing the usual hypothesis $H_0 : \beta = 0$. Table 4.5 shows empirical size and power for Wald tests of H_0 under scenario N1. The tests achieve the nominal level, most exhibiting some elevation when $n = 100$. The proposed approach yields 10–25% increases in power over the nearest competitors except for $n = 500$ with large alternatives.

4.2 Missing at random follow-up

We carried out several simulation studies to assess performance of the proposed approach. We report here on results for a quadratic scenario involving 1000 Monte Carlo replications, $\beta = 0.5$, $\delta = 0.5$, $Y_1 \sim \mathcal{N}(0, 1)$ ($\mu_1 = 0$, $\sigma_{11} = 1$) and intermediate covariate X_1 with $X_1|Y_1, Z \sim \text{Ber}(p_{X_1} = (1 - Z)\texttt{exp}(\kappa_0^{(0)} + \kappa_1^{(0)}Y_1)/\{1 + \texttt{exp}(\kappa_0^{(0)} + \kappa_1^{(0)}Y_1)\} + Z\texttt{exp}(\kappa_0^{(1)} + \kappa_1^{(1)}Y_1)/\{1 + \texttt{exp}(\kappa_0^{(1)} + \kappa_1^{(1)}Y_1)\})$ with parameter values set to $(\kappa_0^{(0)}, \kappa_1^{(0)}) = (1, -0.1)$ and $(\kappa_0^{(1)}, \kappa_1^{(1)}) = (-0.5, 0.1)$. The missing data generating mechanism was simulated via logistic regression modeling with probability $\pi^{(c)}$ equal to $\pi^{(c)} = (1-Z)\texttt{exp}(\gamma_0^{(0)}+\gamma_1^{(0)}Y_1+\gamma_2^{(0)}X_1+\gamma_3^{(0)}X_1Y_1)/\{1+\texttt{exp}(\gamma_0^{(0)}+\gamma_1^{(0)}Y_1+\gamma_2^{(0)}X_1+\gamma_3^{(0)}X_1Y_1)\}+Z\texttt{exp}(\gamma_0^{(1)}+\gamma_1^{(1)}Y_1+\gamma_2^{(1)}X_1+\gamma_3^{(1)}X_1Y_1)/\{1+\texttt{exp}(\gamma_0^{(1)}+\gamma_1^{(1)}Y_1+\gamma_2^{(1)}X_1+\gamma_3^{(1)}X_1Y_1)\}$ with parameter values set to $(\gamma_0^{(0)}, \gamma_1^{(0)}, \gamma_2^{(0)}, \gamma_3^{(0)}) = (0.2, 2.0, 0.1, 0.1)$ and

$(\gamma_0^{(1)}, \gamma_1^{(1)}, \gamma_2^{(1)}, \gamma_3^{(1)}) = (1.0, -1.0, -0.1, 0.1)$; resulting in an approximated proportion of complete cases of 63% and 70% in the control and treatment groups, respectively. The parameter of interest, β, was estimated, using complete cases only, by the paired t-test, ANCOVA I and ANCOVA II and, using all cases, with IWCC and the proposed estimator with linear and/or quadratic polynomial basis for Y_1 used as the polynomial basis $f(Y_1, X_0) = f(Y_1)$ and linear and/or quadratic polynomial function of Y_1 and a linear polynomial function of X_1 as the polynomial basis $\ell(Y_1, X_0, X_1) = \ell(Y_1, X_1)$ for both treatment and control groups, with probability of missingness correctly estimated from a logistic regression with the covariates Y_1, X_1 and their interaction Y_1X_1, thus $d(Y_1, X_1) = (1, Y_1, X_1, Y_1X_1)^T$ for both treatment groups; or computing the sample proportion of follow-up complete cases in each treatment group as an incorrectly specified model for the missing data generating mechanism. Note a quadratic polynomial basis specification for X_1 is redundant since X_1 is a bernoulli random variable. Thus, QUADQUAD_LR estimator has basis functions $f(Y_1) = (1, Y_1, Y_1^2)^T$ and $\ell(Y_1, X_1) = (Y_1^2)$ and BASIS_LR has functions $f(Y_1) = (1Y_1Y_1^2)^T$ and $\ell(Y_1, X_1) = (Y_1^2, Y_1^2X_1)^T$. We also included estimators discussed at the beginning of section 3.2, based on the efficient influence function given in formulæ in 2.2.15 referred to as LOESS and REG, where $E(Y_2|Y_1, Z = c)$ and $E(Y_2|Y_1, X_1, Z = c)$, $c = 0, 1$ were estimated via local polynomial smoothing using `proc loess` in SAS (SAS Institute, 2000) or via linear regression respectively, including an intercept and the covariates $(Y_1, Y_1^2)^T$ in modeling $E(Y_2|Y_1, Z = c)$, $c = 0, 1$ and the covariates $(1, Y_1, X_1, Y_1X_1, Y_1^2, Y_1^2X_1)^T$ for modeling $E(Y_2|Y_1, X_1, Z = c)$, $c = 0, 1$,

substituting in 2.2.15 or(2.2.10) and analog, finding estimators for $\mu_2^{(0)}$ and $\mu_2^{(1)}$ by equating sample averages of each term in (2.2.10) to zero, and taking their difference. A logistic model was used in both estimators to estimate the missing data mechanism. To verify the robustness property discussed at the end of section 2.2, an estimator called REG was included, with same modeling specifactions as REG_LR but with missingness estimated using sample proportions. For the Paired t-test, ANCOVA I and ANCOVA II estimators, standard error estimates were obtained as in the full-data case in 4.1. For IWCC_LR, BASIS_LR, QUADQUAD_LR, standard errors were obtained from (3.2.10) and a sandwich approach using $(\sum_{i=1}^{n} S_i^2)^{1/2}$, where S_i is the difference of the corresponding influence function of the estimator in (3.2.9) and its analog for the control mean $\mu_2^{(0)}$, for each i with all unknown quantities replaced by sample analogs. For , LOESS and REG, standard errors were obtained from 3.2.7, and a sandwich approach using S_i from 2.2.15, for each i with all unknown quantities replaced by sample analogs as well. Nominal 95% Wald confidence intervals for β were constructed as the estimate ± 1.96 times the asymptotic and sandwich formulæ standard error estimates.

Follow-up responses for the quadratic scenario were generated using the quadratic model

$$Y_{2i} = (\mu_2 + \beta Z_i) + \{\beta_1 + \beta_2(X_{1i} - E(X_{1i}|Z))\}(Y_{1i} - \mu_1) + \{\beta_3 + \beta_4(X_{1i} - E(X_{1i}|Z))\}\{(Y_{1i} - \mu_1)^2 - \sigma_{11}\} + \beta_5(X_{1i} - E(X_{1i}|Z) + \epsilon_i, (4.2.3)$$

where $\epsilon_i \sim \mathcal{N}((0,1)$, considered by Yang and Tsiatis (2001), with $\mu_2 = 0$. Normality

of baseline and follow-up responses may be considered the "most favorable" distribution for the paired t-test and the ANCOVA II estimators, which are often thought to be predicated on normality, although the results of Yang and Tsiatis and Leon et. al. (2003) show that these estimators are consistent under general conditions.

Situation Q1 is based on (4.2.3) with $(\beta_1, \beta_2, \beta_3, \beta_4, \beta_5) = (0.5, 0.3, 0.4, 0.3, 0.4)$, corresponding to a discernible curvilinear relationship between baseline and follow-up, as depicted in Figure 4.2(a). the paired t-test and ANCOVA II estimators assume a linear relationship, thus, it is of interest to note the linear correlations of 0.62 and 0.61 between baseline and follow-up for the control and treatment groups, respectively. The second situation, Q2, exemplified in Figure 4.2(b), takes $(\beta_1, \beta_2, \beta_3, \beta_4, \beta_5) = (0.1, -0.1, 0.2, -0.1, 0.1)$ and results in "low" correlations of 0.32 and 0.33 and same percent missingness as situation Q1. Tables 4.6 and 4.7 shows the results for $n = 1000$.

The BENCHMARK estimator captures the true form of both $E(Y_2|Y_1, Z)$ and $E(Y_2|Y_1, X_1, Z)$.

For scenario Q1, the estimators based on complete cases such as paired t-test, ANCOVA I and ANCOVA II showed relative biases over 10%, expect the paired t-test, who showed a relative bias of 7%; also efficiency was degraded by almost half in comparison to the BENCHMARK estimator. All estimators that attempted to model missingness showed relative biases under 2% and high relative efficiency in comparison to BENCHMARK. Note REG and IWCC_LR show the "doubly robustness property" as is indicated in the results by having almost identical values shown in 4.6 for the

REG_LR and BENCHMARK estimators that correctly specify (or estimate) a model for $E(Y_2|Y_1, X_1, Z)$ and missingness. REG misspecifies a model for missingness and IWCC_LR mispecifies a model for $E(Y_2|Y_1, X_1, Z)$. The degradation in efficiency that might be expected from missspecifying a correct model for $E(Y_2|Y_1, Z)$ in the REG estimator was not observed in 4.6.

In the case of scenario Q2, all estimators perform the same as expected. However, a relative bias over 4% is observed in estimators based on complete cases. The same efficiency is acheived by all estimators.

These simulations and others not reported suggest that, in general, larger biases are introduced in estimators based on complete cases when follow-up observations are missing at random. Estimators based on the efficient function or on the basis functions approach improved efficiency, while they showed very small relative bias; which suggest they are a plausible alternative for estimating treatment effect under conditions highlighted along the dissertation.

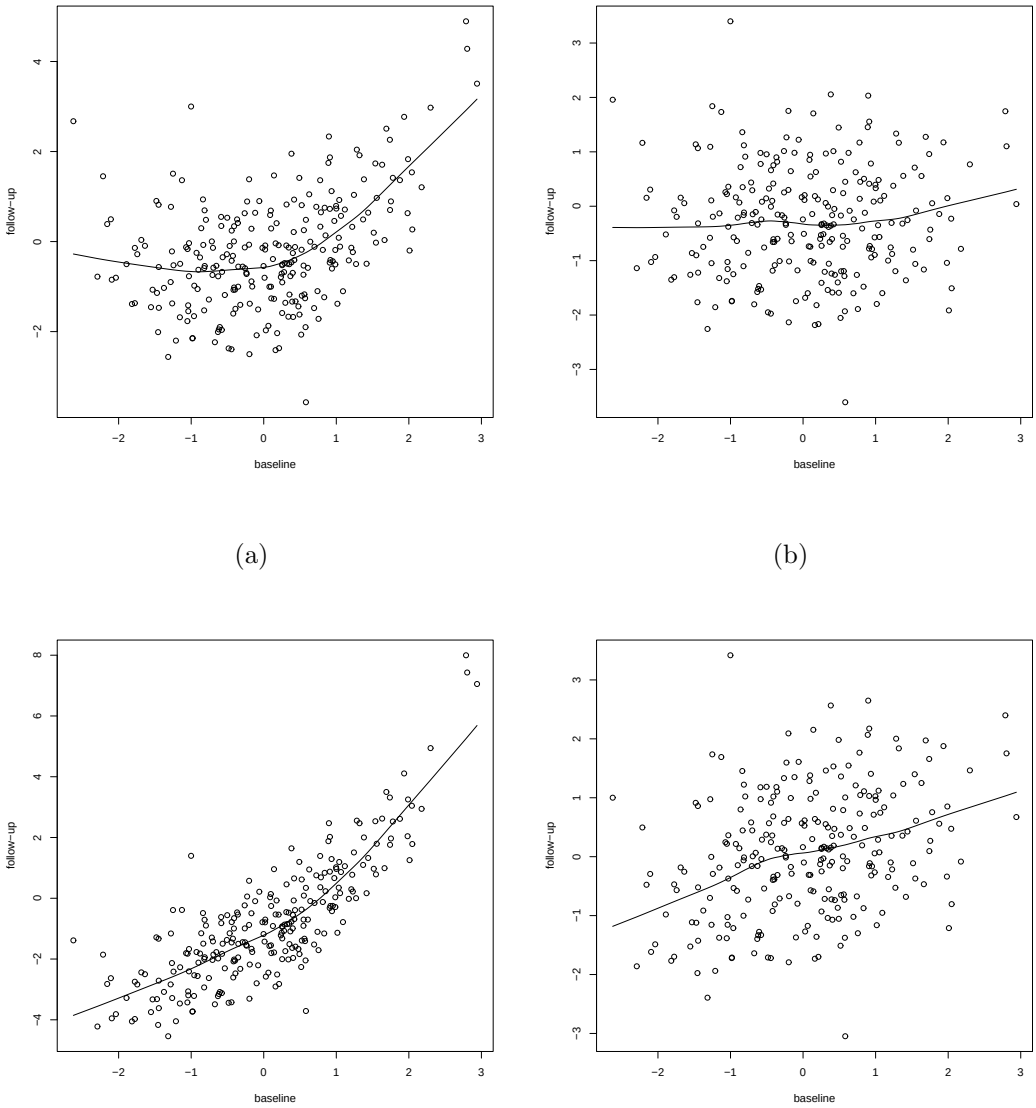


Figure 4.1: Simulated data for $n = 500$ from scenarios **a.** Q1 and **b.** based on (4.1.1) and scenarios **c.** N1 and **d.** N2 based on (4.1.2), with smooth fits (solid line) obtained using the Splus function `lowess()` (Cleveland, 1979). Each panel depicts data for roughly 250 subjects randomized to control.

Table 4.1: Simulation results for true quadratic relationship (4.1.1), 5000 Monte Carlo data sets. Estimators and scenarios are as described in the text. MC Mean is Monte Carlo average, MC SD is Monte Carlo standard deviation, Asymp. SE is the average of estimated standard errors based on the asymptotic theory, OLS SE is the average of estimated standard errors based on OLS for the "popular" estimators, MSE Ratio is Mean Square Error (MSE) for QUAD divided by MSE of the indicated estimator, CP is empirical coverage probability of confidence interval using asymptotic SEs.

Estimator	MC Mean	MC SD	Asymp. SE	OLS SE	MSE Ratio	CP
			Scenario Q1, "moderate" association			
			n=100			
LOESS	0.504	0.218	0.202	–	0.89	0.93
QUAD	0.507	0.205	0.200	–	1.00	0.93
ANCOVA II	0.506	0.237	0.228	0.231	0.75	0.94
ANCOVA I	0.506	0.233	0.228	0.231	0.77	0.94
Paired t-test	0.508	0.255	0.251	0.251	0.65	0.95
Two sample t-test	0.506	0.254	0.251	0.251	0.65	0.94
			n=500			
LOESS	0.501	0.090	0.091	–	0.98	0.96
QUAD	0.501	0.089	0.089	–	1.00	0.95
ANCOVA II	0.502	0.103	0.103	0.103	0.75	0.95
ANCOVA I	0.502	0.102	0.103	0.103	0.76	0.95
Paired t-test	0.501	0.111	0.112	0.112	0.64	0.95
Two sample t-test	0.503	0.112	0.112	0.112	0.63	0.95

Table 4.2: Simulation results for true quadratic relationship (4.1.1), 5000 Monte Carlo data sets. Columns are as described in table 4.1. Estimators and scenarios are as described in the text.

Estimator	MC Mean	MC SD	Asymp. SE	OLS SE	MSE Ratio	CP
			Scenario Q2, "weak" association			
			n=100			
LOESS	0.506	0.214	0.189	–	0.90	0.91
QUAD	0.505	0.203	0.199	–	1.00	0.94
ANCOVA II	0.505	0.205	0.202	0.205	0.98	0.94
ANCOVA I	0.505	0.204	0.202	0.204	0.99	0.94
Paired t-test	0.511	0.269	0.272	0.272	0.57	0.95
Two sample t-test	0.505	0.204	0.204	0.204	0.99	0.95
			n=500			
LOESS	0.499	0.091	0.089	–	0.97	0.94
QUAD	0.499	0.089	0.089	–	1.00	0.95
ANCOVA II	0.499	0.091	0.090	0.091	0.98	0.95
ANCOVA I	0.499	0.090	0.090	0.090	0.98	0.95
Paired t-test	0.499	0.121	0.121	0.121	0.55	0.95
Two sample t-test	0.499	0.091	0.091	0.091	0.97	0.95

Table 4.3: Simulation results for true nonlinear relationship (4.1.2), 5000 Monte Carlo data sets. Estimators and scenarios are as described in the text. Column headings are as in Table 4.1.

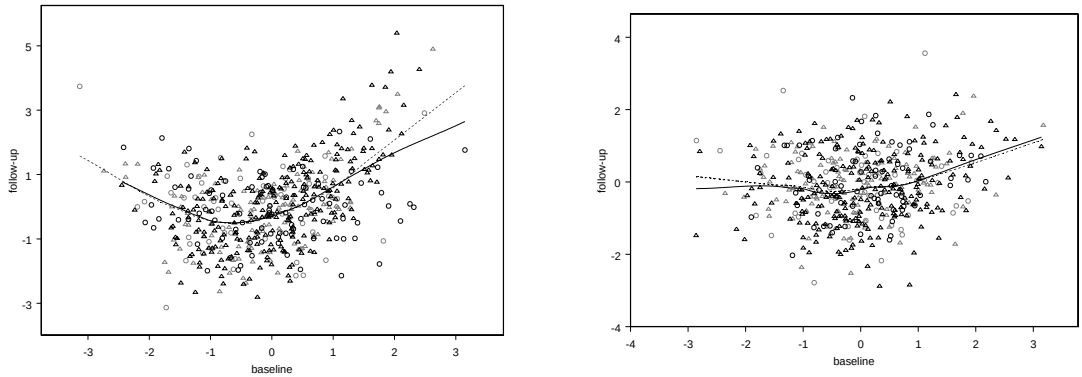
Estimator	MC Mean	MC SD	Asymp. SE	OLS SE	MSE Ratio	CP
			Scenario N1, "high" association			
			n=100			
BENCHMARK	0.507	0.203	0.205	–	1.05	0.94
LOESS	0.504	0.218	0.237	–	0.91	0.96
QUAD	0.507	0.207	0.201	–	1.00	0.93
ANCOVA II	0.506	0.236	0.228	0.231	0.77	0.93
ANCOVA I	0.506	0.232	0.228	0.231	0.80	0.94
Paired t-test	0.505	0.257	0.254	0.254	0.65	0.94
Two sample t-test	0.503	0.387	0.384	0.384	0.29	0.94
			n=500			
BENCHMARK	0.501	0.089	0.092	–	1.03	0.96
LOESS	0.501	0.090	0.096	–	1.00	0.96
QUAD	0.501	0.090	0.090	–	1.00	0.95
ANCOVA II	0.502	0.103	0.103	0.103	0.77	0.95
ANCOVA I	0.502	0.103	0.103	0.103	0.77	0.95
Paired t-test	0.502	0.114	0.114	0.114	0.63	0.95
Two sample t-test	0.504	0.173	0.172	0.172	0.27	0.95

Table 4.4: Simulation results for true nonlinear relationship (4.1.2), 5000 Monte Carlo data sets. Estimators and scenarios are as described in the text. Column headings are as in Table 4.1.

Estimator	MC Mean	MC SD	Asymp. SE	OLS SE	MSE Ratio	CP
			Scenario N2, "mild" association			
			n = 100			
BENCHMARK	0.505	0.201	0.204	–	1.03	0.95
LOESS	0.506	0.214	0.191	–	0.90	0.92
QUAD	0.505	0.203	0.199	–	1.00	0.94
ANCOVA II	0.505	0.203	0.200	0.203	1.00	0.94
ANCOVA I	0.505	0.202	0.200	0.202	1.01	0.94
Paired t-test	0.509	0.231	0.233	0.233	0.77	0.95
Two sample t-test	0.503	0.219	0.217	0.217	0.86	0.95
			n = 500			
BENCHMARK	0.499	0.089	0.091	–	1.00	0.96
LOESS	0.499	0.091	0.089	–	0.97	0.95
QUAD	0.499	0.089	0.089	–	1.00	0.95
ANCOVA II	0.499	0.090	0.089	0.090	1.00	0.95
ANCOVA I	0.499	0.089	0.089	0.090	1.00	0.95
Paired t-test	0.499	0.103	0.104	0.104	0.75	0.95
Two sample t-test	0.499	0.098	0.097	0.097	0.84	0.95

Table 4.5: Empirical size and power of Wald tests (estimate/asymptotic standard error estimate) of $H_0 : \beta = 0$ under scenario N1 with (4.1.2), each based on 5000 Monte Carlo data sets. Empirical size was found by simulations with $\beta = 0$. Empirical power is under the indicated alternative.

	Size		Power					
	$\beta = 0$		$\beta = 0.25$		$\beta = 0.40$		$\beta = 0.50$	
n	100	500	100	500	100	500	100	500
BENCHMARK	0.07	0.05	0.26	0.80	0.52	0.99	0.70	1.00
LOESS	0.04	0.04	0.18	0.76	0.39	0.99	0.56	1.00
QUAD	0.07	0.05	0.25	0.79	0.53	0.99	0.71	1.00
ANCOVA II	0.07	0.05	0.21	0.69	0.43	0.97	0.60	1.00
ANCOVA I	0.06	0.05	0.21	0.69	0.43	0.97	0.60	1.00
Paired t-test	0.06	0.05	0.17	0.61	0.37	0.94	0.52	0.99
Two sample t-test	0.06	0.05	0.11	0.32	0.19	0.66	0.26	0.83



(a) (b)

Figure 4.2: Simulated data for $n = 1000$ from scenarios **a.** Q1 and **b.** based on (4.2.3) and scenarios **c.**, with smooth fits obtained using the Splus function `loess()` (Cleveland, 1979). Solid lines are fitted utilizing only complete cases (observations in black), in contrast, dotted lines are based on the "full" data, as if all observations were available (cases considered as missing follow-up response are colored in light grey). Data represented with triangles are observations with intermediate bernoulli value $X_1 = 1$. Each panel depicts data for roughly 500 subjects randomized to control.

Table 4.6: Simulation results for true quadratic relationship (4.2.3), 1000 Monte Carlo data sets with sample size of $n = 1000$. Estimators and scenarios are as described in the text. Bias is the percent relative bias. Columns headings are as described in table 4.1. MSE Ratio is Mean Square Error (MSE) for BENCHMARK divided by MSE of the indicated estimator, CP and CPsw are empirical coverage probabilities of confidence interval using asymptotic and sandwich SEs respectively. Treatment effect parameter value is $\beta = 0.51$.

	MC		Standard Error				MSE		
Estimator	Mean	Bias	MC SD	Asymp.	OLS	Sand.	ratio	CP	CPsw
			Scenario Q1, "moderate" association						
BENCHMARK	0.50	-1.8	0.08	0.07	–	0.08	1.00	0.94	0.95
LOESS_LR	0.51	0.1	0.08	0.08	–	0.08	0.98	0.95	0.95
REG_LR	0.52	0.9	0.08	0.07	–	0.08	1.00	0.94	0.95
BASIS_LR	0.51	-0.4	0.08	0.08	–	0.08	0.97	0.95	0.96
QUADQUAD_LR	0.51	0.0	0.08	0.08	–	0.08	0.97	0.95	0.96
IWCC_LR	0.52	1.1	0.10	0.10	–	0.10	0.85	0.85	0.85
REG	0.52	0.9	0.08	0.08	–	0.08	1.00	0.95	0.95
IWCC	0.43	-16.3	0.09	0.10	–	0.10	0.38	0.85	0.85
GEE	0.45	-11.6	0.09	0.09	–	–	0.55	0.90	–
ANCOVA II	0.45	-11.6	0.09	0.09	0.09	–	0.55	0.90	–
ANCOVA I	0.45	-11.5	0.09	0.09	0.09	–	0.56	0.90	–
Paired t-test	0.48	-6.9	0.09	0.10	0.10	–	0.61	0.94	–

Table 4.7: Simulation results for true quadratic relationship (4.2.3), 1000 Monte Carlo data sets. Estimators and scenarios are as described in the text. Column headings are as decribe in Table 4.6. The total sample size is $n = 1000$ and treatment effect parameter value is $\beta = 0.50$.

	MC		Standard Error				MSE		
Estimator	Mean	Bias	MC SD	Asymp.	OLS	Sand.	ratio	CP	CPsw
Scenario Q2, "weak" association									
BENCHMARK	0.50	1.1	0.08	0.07	–	0.08	1.00	0.92	0.93
LOESS_LR	0.50	-0.3	0.08	0.07	–	0.07	0.99	0.92	0.93
REG_LR	0.50	0.2	0.08	0.07	–	0.08	1.00	0.93	0.93
BASIS_LR	0.50	0.1	0.08	0.08	–	0.08	0.99	0.93	0.93
QUADQUAD_LR	0.50	0.3	0.08	0.08	–	0.08	1.00	0.93	0.94
IWCC_LR	0.50	0.3	0.08	0.08	–	0.08	0.99	0.93	0.94
REG	0.50	0.2	0.08	0.07	–	0.07	1.00	0.93	0.93
IWCC	0.50	-4.2	0.08	0.08	–	0.08	0.90	0.93	0.93
GEE	0.48	-3.5	0.08	0.08	–	–	0.93	0.93	–
ANCOVA II	0.48	-3.5	0.08	0.08	0.08	–	0.92	0.93	–
ANCOVA I	0.48	-3.4	0.08	0.08	0.08	–	0.93	0.93	–
Paired t-test	0.52	4.6	0.10	0.10	0.10	–	0.53	0.94	–

Chapter 5

Treatment effect in ACTG 175

We apply methods developed in previous chapters to evaluate treatment effect in CD4 count at 20 ± 5 and 96 ± 5 week post-randomization.

5.1 20 ± 5 CD4

Figure 1.1 shows a possibly nonlinear baseline-follow-up relationship in each ACTG 175 group (with correlations of 0.55 and 0.65 for control and treatment); here, $\delta = 0.75$. It is standard to analyze CD4 counts on a scale where they appear symmetrically distributed to achieve the approximate normality widely thought required for valid inference via "popular" methods, although Section 2.1 shows this is not needed for consistency. If interest focuses on mean CD4, inference under a monotone transformation to normality instead addresses median CD4.

Table 5.1 presents results using the approaches studied in Section 4 for data from

2130 subjects, excluding those missing baseline or 20 ± 5 week CD4. Also given are two estimators incorporating baseline covariates $X_0 = \{$weight (kg), Karnofsky score (0-100), days of pre-study antiretroviral therapy, symptomatic status (0/1), IV drug use (0/1)$\}$, one using basis functions $f(X_0, Y_1) = (1, Y_1, Y_1^2, X_{01}, \ldots, X_{05})^T$ and the other obtained by fitting $E(Y_2|X_0, Y_1, Z = c)$, $c = 0, 1$, via generalized additive models (Hastie and Tibshirani, 1990). The proposed methods have the smallest standard errors, and incorporation of baseline covariates yields further improvement. The gains are slight, however, likely owing to the weak curvature in Figure 1.1 and weak prognostic effect of the covariates.

5.2 96 ± 5 CD4

In many HIV studies, loss to follow-up is often thought a consequence of review by participants and their healthcare providers of available information, such as intermediate measures of disease status; dropout may also be more likely among subjects with certain baseline characteristics. To illustrate the proposed methods, we consider data from AIDS Clinical Trials Group (ACTG) protocol 175, which randomized 2467 HIV-infected individuals to four groups: zidovudine (ZDV) monotherapy, ZDV plus didanosine (ddI), ZDV plus zalcitabine (ddC), or ddI (Hammer et al. 1996). On the basis of the primary outcome, time to progression to AIDS or death, ZDV was shown inferior to the other three treatments. Thus, we consider two interventions, ZDV alone ("control") and the combination of the other three ("treatment"). Secondary analy-

ses in such trials often focus on changes in measures of immune or virologic response; accordingly, we consider follow-up CD4 count at 96 ± 5 weeks post-randomization, missing for 37% of the subjects. Figure 1.2 shows the relationship between baseline and follow-up CD4 among subjects for whom both are available and suggests a positive, possibly curvilinear relationship in each group. In addition to baseline CD4, numerous covariates, such as hemophilia status, sexual preference, intravenous drug use, Karnofsky score, prior antiretroviral treatment experience, gender, race, and so on, were recorded at baseline, and intervening measures of CD4, CD8, and treatment status were collected over the 96-week period. Exploratory analyses of associations between these covariates and follow-up status show that missingness is related to a several of them, in many cases differentially by treatment.

Figure 1.2 shows that 96 ± 5 CD4 follows a similar trend with respect to baseline CD4 in each ACTG 175 group (with correlations of 0.55 and 0.54 for control and treatment).

Table 5.2 presents results using the approaches studied in Section 4 for data from 2120 subjects, excluding those cases missing baseline or intervening covariate. Paired t-test, ANCOVA I and ANCOVA II were estimated from 1460 cases with available 96 ± 5 week CD4. Also given are two estimators incorporating baseline covariates $X_0 = \{$weight (kg), Karnofsky score (0-100), days of pre-study antiretroviral therapy, symptomatic status (0/1), IV drug use (0/1)$\}$ and follow-up covariates $X_1 = \{20 \pm 5$ week CD4, 20 ± 5 week CD8$\}$, one using basis functions $f(X_0, Y_1) = (1, Y_1, Y_1^2, X_{01}, \ldots, X_{05})^T$ and $\ell(X_1, X_0, Y_1) = (1, Y_1^2, X_{11}^2, X_{12}^2)^T$ and

modelling the probability of missingness with a logistic regression with all baseline covariates and linear polynomials of the follow-up covariates. The estimator called REG, obtained by fitting $E(Y_2|X_0, Y_1, Z = c)$, $c = 0, 1$, via linear regression on the same covariates as the basis function $f(Y_1, X_0)$ and fitting $E(Y_2|X_0, Y_1, X_1, Z = c)$, $c = 0, 1$, via linear regression on the same covariates as the basis function $\ell(Y_1, X_0, X_1)$ is also given, with missingness modeled using sample proportions.

The proposed methods have the smallest standard errors, and incorporation of baseline covariates yields further improvement. We see a potential bias introduced in the estimation of treatment effect in ANCOVA I, ANCOVA II and paired t-test estimators. However, both bias and improvement in the standard error estimation is slight, probably because of the weak curvature showed in Figure 1.2 and weak prognostic factor effect of the covariates.

Table 5.1: Treatment effect estimates for 20 ± 5 week CD4 counts for ACTG 175. The methods are as denoted as in Tables 4.1-4.5; in addition, BASE-QUAD denotes the proposed basis function method including up to quadratic terms in baseline CD4 and linear terms involving baseline covariates, and BASE-GAM denotes the proposed method estimating the conditional expectations using generalized additive models. Asymptotic SE is estimated standard error based on the influence function and OLS SE is estimated standard error based on the "usual" approaches for the "popular" estimators as described in Section 4.

Estimator	$\widehat{\beta}$	Asymptotic SE	OLS SE
BASE-GAM	50.001	5.080	–
BASE-QUAD	50.837	4.957	–
LOESS	49.799	5.222	–
QUAD	49.792	5.330	–
ANCOVA II	49.404	5.384	5.842
ANCOVA I	49.313	5.385	5.840
Paired t-test	50.148	5.686	6.109
Two sample t-test	45.506	6.767	7.203

Table 5.2: Treatment effect estimates for 96 ± 5 week CD4 counts for ACTG 175. BASIS_LR denotes the proposed approximating method using quadratic polynomial basis and intermediate covariates for modeling missingness. Asymp. SE is estimated standard error based on asymptotic formulæ and OLS SE is estimated standard error based on the "usual" approaches for the "popular" estimators as described in Chapter 4.

Estimator	$\widehat{\beta}$	Asymp. SE	OLS SE	Sandwich SE
REG_LR	62.87	7.91	–	8.22
BASIS_LR	63.13	8.05	–	8.31
REG	63.32	8.19	–	8.33
ANCOVA I	65.66	8.27	8.75	–
ANCOVA II	65.59	8.45	8.80	–
Paired t-test	67.57	8.94	8.93	–
Two sample t-test	54.50	10.36	10.53	–

Chapter 6

Discussion

Casting the pretest-posttest problem in a counterfactual framework, we have exploited advances in the missing data literature to characterize a general class of estimators for treatment effect. We have shown that "popular" approaches are inefficient and that improvement is possible by taking into account the nature of the relationship between baseline and follow-up response via our approach. We do not recommend the proposed estimators in very small samples, as they require estimation of features that may be not be well-identified under these conditions. The formulation also yields an approach for incorporating baseline covariate information to further improve efficiency. Our strategy differs from that of directly modeling parametrically the relationship between follow-up and baseline and other factors. In this approach, consistency is predicated on correct specification of the model (and possible distributional assumptions), while that here will yield consistent estimators of treatment effect regardless, e.g., if the chosen basis does not accurately reflect the true relationship.

When there exists an interaction between baseline response and treatment, a philosophical issue is whether interest should indeed focus on main effects. Proponents of the "large, simple trial" (e.g., Friedman, Furberg, and DeMets, 1996, p. 56) downplay interactions and focus on overall effect, while researchers in other settings have a different view. We do not take a position in this debate. The proposed estimators are for such a main effect; whether there is an interaction or not, the methods estimate this effect consistently and exploit relationships among variables solely to gain efficiency.

It is in fact straightforward to extend the development to more than two treatments and to observational studies where treatment assignment is not random and it is assumed that treatment assignment is independent of prognosis (i.e., follow-up counterfactuals) given baseline covariates.

We extended the method to include cases with follow-up missing data, and showed the utility of the proposed estimators when the data is missing at random. We see that meaningful bias can be introduced in the estimation if failed to considering the possible underlying missing mechanism present in the data. We believe that the contribution of Robins et. al. (1994) to modeling missing data can be best assessed by making such methods easily available on settings commonly found by practitioners, like the pretest-posttest problem. Also, the simplicity of the problem allows to elucidate understanding about the utility of those methods for modeling missing data.

Bibliography

Bickel, P., K. C. R. Y. & Wellner, J. (1993). Efficient and adaptive inference in semiparametric models. *Baltimore: Johns Hopkins University Press*.

Cleveland, W. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association* **74**, 829–836.

Crager, M. R. (1980). Analysis of covariance in parallel group clinical trials with pretreatment baseline. *Biometrics* **43**, 895–901.

der Laan, V. & Robins, J. (2003). *Unified Methods for Censored Longitudinal Data and Causality*. New York: Springer.

Follmann, D. A. (1991). The effect of screening on some pretest-posttest test variances. *Biometrics* **47**, 763–771.

Friedman, L.M., F. C. & DeMets, D. (1996). *Fundamentals of Clinical Trials*. 3rd edn. St. Louis: Mosby.

Hammer, S. M., Katezstein, D. A., Hughes, M. D., Gundaker, H., Schooley, B. T., Haubrich, R. H., Henry, W. K., Lederman, M. M. & Phair, J. P. (1996). A trial

comparing nucleoside monotherapy with combination therapy in hiv-infected adults with cd4 cell counts from 200 to 500 per cubic millimiter. *New England Journal of Medicine* **335**, 1081–1089.

Hastie, T. & Tibshirani, R. (1990). *Generalized Additive Models*. London: Chapman and Hall.

Horvitz, D. G. & Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association* **47**, 663–685.

Institute, S. (2000). *SAS/STAT User's Guide. Version 8*. 4th edn. Cary, North Carolina: SAS Publishing.

Laird, N. (1983). Further comparative analyses of pretest-posttest research designs. *The American Statistician* **37**, 329–340.

Newey, W. K. (1990). Semiparametric efficiency bounds. *Journal of Applied Econometrics* **5**, 99–135.

Robins, J. M. (1999). Robust estimation in sequentially ignorable missing data and causal inference models. *Proceedings of the American Statistical Association Section on Bayesian Statistical Science* pp. 6–10.

Singer, J. M. & Andrade, D. F. (1997). Regression models for the analysis of pretest/posttest data. *Biological Psychology* **53**, 729–735.

Stanek, E. J. I. (1988). Choosing a pretest-posttest analysis. *The American Statistician* **42**, 178–183.

Stein, R. A. (1989). Adjusting treatment effects for baseline and other predictor variables. *ASA Biopharmaceutical Section* pp. 274–280.

Yang, L. & Tsiatis, A. A. (2001). Efficiency study of estimators for treatment effect in a pretest-posttest trial. *The American Statistician* **55**, 314–321.