

ABSTRACT

CAO, WEIHUA. Improving Efficiency and Robustness of Doubly Robust Estimators in the Presence of Coarsened Data. (Under the direction of Dr. Marie Davidian and Dr. Anastasios A. Tsiatis).

Considerable recent interest has focused on doubly robust estimators for a population mean response in the presence of incomplete data, which involve models for both the propensity score and the regression of outcome on covariates. The “usual” doubly robust estimator may yield severely biased inferences if neither of these models is correctly specified and can exhibit nonnegligible bias if the estimated propensity score is close to zero for some observations. In part one of this dissertation, we propose alternative doubly robust estimators that achieve comparable or improved performance relative to existing methods, even with some estimated propensity scores close to zero.

The second part of this dissertation focuses on drawing inference on parameters in general models in the presence of monotonely coarsened data, which can be viewed as a generalization of longitudinal data with a monotone missingness pattern, as is the case when subjects drop out of a study. Estimators for parameters of interest include both inverse probability weighted estimators and doubly robust estimators. As a generalization of methods in part one, we propose alternative doubly robust estimators that achieve comparable or improved performance relative to existing methods. We apply the proposed method to data from an AIDS clinical trial.

Improving Efficiency and Robustness of Doubly Robust Estimators in the Presence of
Coarsened Data

by
Weihua Cao

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, North Carolina

2009

APPROVED BY:

Dr. Dennis Boos

Dr. Daowen Zhang

Dr. Marie Davidian
Chair of Advisory Committee

Dr. Anastasios A. Tsiatis
Co-Chair of Advisory Committee

DEDICATION

To my loving family.

BIOGRAPHY

Weihua Cao was born in a small village near Xuchang City, Henan Province, China in 1981. In 1998, for the first time in his life, he stepped out of Henan Province and went to Shanghai to attend Fudan University studying mathematics. Upon obtaining his bachelor degree in 2003, he traveled across the sea to United States for graduate study at Case Western Reserve University, and graduated with a master degree in applied mathematics in 2005. Then he moved to Purdue University to pursue a Ph.D. degree in statistics. In 2006, he transferred to North Carolina State University to continue his study in statistics. He married to Zhi Wen in 2005, and they have a lovely 3 year old daughter Amy Cao. He expects to complete his Ph.D. in December, 2009.

ACKNOWLEDGMENTS

I owe my deepest gratitude to to my advisors, Dr. Marie Davidian and Dr. Anastasios Tsiatis. I feel so fortunate and honored to have you as my advisors. Your encouragement, talented guidance and support are always the driving forces during my Ph.D. study. You not only help me with these specific research problems, but more importantly also show me the right way to do research, which is a vivid description of the old saying: teach a man how to fish feeds for life. I only wish I did not let you know in the past three years.

I would like to express my sincere appreciate to Dr. Dennis Boos and Dr. Daowen Zhang, for their serving on my committee, for their valuable suggestions on this research, and for their help during my job hunting. I also would like to thank their tutoring since I took five great courses from them.

It is an honor for me to be in the last class of Dr. Bibhuti Bhattacharyya before he retired. It never stopped fascinating me that how could he make measure theory so interesting and full of fun. He set a role model in teaching for me. I am thankful to Dr. Pam Arway for her help during my graduate study. I also would like to thank my managers at GlaxoSmithKline, Dr. Mandy Bergquist and Dr. Karen Chiswell, who taught me a lot of things in industry and guided me through my one year graduate industrial trainee over there.

I am deeply indebted to my parents, brothers, wife and daughter for their endless love and support. Special thanks go to my wife Zhi and my daughter Amy, you are the meaning of my life.

Lastly, I offer my regards and blessings to all of those who supported me in any respect during my study and the completion of the thesis.

TABLE OF CONTENTS

LIST OF TABLES.....	vi
1 Improving efficiency and robustness of the doubly robust estimator for a population mean with incomplete data	1
1.1 Introduction	1
1.2 Existing doubly robust estimators	2
1.3 Alternative doubly robust estimators	4
1.4 Enhanced propensity score model	9
1.5 Simulation studies	10
1.6 Discussion	11
1.7 Details	13
1.7.1 Optimal, doubly robust estimator for a treatment mean difference	13
1.7.2 Enhanced propensity score and stabilized weights	19
1.7.3 Calculation of asymptotic variances	20
1.7.4 Asymptotic bias under local misspecification	23
2 Improving Efficiency and Robustness of Doubly Robust Estimators in the Presence of Coarsened Data	27
2.1 Introduction	27
2.2 General coarsened data framework and coarsening at random	30
2.3 Inferential objective and doubly robust estimators	31
2.4 Existing and proposed doubly robust estimators	33
2.5 Application to ACTG 175	39
2.6 Simulation studies	42
2.7 Discussion	45
2.8 Details	46
2.8.1 Derivation of approximate standard errors via the sandwich method	46
2.8.2 Derivation of conditional expectations implied by assumed mixed models in Section 2.5	50
2.8.3 Derivation of conditional expectations implied by assumed mixed model in Section 2.6	52
Bibliography	54
Appendices.....	57
SAS code for Chapter 1	58
SAS code for Chapter 2	72

LIST OF TABLES

Table 1.1 Simulation results based on 1000 Monte Carlo replications for the Kang and Schafer scenario. Sample size is 200. Bias is Monte Carlo bias, RMSE is root mean square error, MAE is median of absolute errors, MCSD is Monte Carlo standard deviation, AveSE is average of sandwich standard errors, Cov is Monte Carlo coverage of 95% Wald confidence intervals, OR is outcome regression, and PS is propensity score. Smallest, median, second largest, and largest standard errors for entries in Table 1.1, 1.2 : BIAS,(0.04, 0.08, 0.39, 5.58); AVESE, (0.0008, 0.004, 0.58, 6.64); COV, (0.006, 0.007, 0.015, 0.015).	14
Table 1.2 Simulation results based on 1000 Monte Carlo replications for the Kang and Schafer scenario. Sample size is 1000. Bias is Monte Carlo bias, RMSE is root mean square error, MAE is median of absolute errors, MCSD is Monte Carlo standard deviation, AveSE is average of sandwich standard errors, Cov is Monte Carlo coverage of 95% Wald confidence intervals, OR is outcome regression, and PS is propensity score. Smallest, median, second largest, and largest standard errors for entries in Table 1.1, 1.2 : BIAS, (0.04, 0.08, 0.39, 5.58); AVESE, (0.0008, 0.004, 0.58, 6.64); COV, (0.006, 0.007, 0.015, 0.015).	15
Table 1.3 Simulation results based on 1000 Monte Carlo replications for the Tan scenario. Sample size is 200. Entries are as in Table 1.1. The Tan and Kang and Schafer scenarios are distributionally identical in the “OR correct, PS correct” case. Smallest, median, second largest, and largest standard errors for entries in Table 1.3, 1.4: BIAS, (0.04, 0.08, 0.76, 5.66); AVESE, (0.0008,0.004, 1.59, 9.78); COV, (0.006, 0.007, 0.010, 0.015).	16
Table 1.4 Simulation results based on 1000 Monte Carlo replications for the Tan scenario. Sample size is 1000. Entries are as in Table 1.1. The Tan and Kang and Schafer scenarios are distributionally identical in the “OR correct, PS correct” case. Smallest, median, second largest, and largest standard errors for entries in Table 1.3, 1.4: BIAS, (0.04, 0.08, 0.76, 5.66); AVESE, (0.0008,0.004, 1.59, 9.78); COV, (0.006, 0.007, 0.010, 0.015).	17
Table 2.1 Simulation results based on 1000 Monte Carlo replications. Bias is Monte Carlo bias, RMSE is root mean square error, MCSD is Monte Carlo standard deviation, AveSE is average of sandwich standard errors, Cov is Monte Carlo coverage of 95% Wald confidence intervals, R denotes regression models, and DH denotes discrete hazard models. True value of $\beta = 10.5$. Smallest, median, second largest, and largest standard errors for entries in	

Table1 2.1, 2.2: Bias, (0.019, 0.033, 0.067, 0.077); AveSE, (0.004, 0.022, 0.086, 0.340); Cov, (0.007, 0.008, 0.010, 0.011).	44
---	----

Table 2.2 Simulation results based on 1000 Monte Carlo replications. Bias is Monte Carlo bias, RMSE is root mean square error, MCSD is Monte Carlo standard deviation, AveSE is average of sandwich standard errors, Cov is Monte Carlo coverage of 95% Wald confidence intervals, R denotes regression models, and DH denotes discrete hazard models. True value of $\beta = 10.5$. Smallest, median, second largest, and largest standard errors for entries in Table1 2.1, 2.2: Bias, (0.019, 0.033, 0.067, 0.077); AveSE, (0.004, 0.022, 0.086, 0.340); Cov, (0.007, 0.008, 0.010, 0.011).	45
---	----

Chapter 1

Improving efficiency and robustness of the doubly robust estimator for a population mean with incomplete data

1.1 Introduction

The challenge of estimating a population mean response on the basis of incomplete data arises in many settings. Nonresponse in sample surveys or dropout and noncompliance in clinical trials may lead to missing outcomes for some subjects; likewise, making causal inference on a treatment mean may be viewed as a missing data problem, where potential outcomes under treatment are “missing” for subjects actually observed to receive control (Kang and Schafer, 2007). In these situations, unless the missingness mechanism is “completely at random” (Rubin, 1976), it is well known that the naive sample mean based on the “complete cases” is a biased estimator.

If missing data can reasonably be assumed missing at random, or, equivalently, if the no unmeasured confounders assumption (Rosenbaum and Rubin, 1983; Robins, Hernán, and Brumback, 2000) is tenable when making causal inference from observational data, popular approaches include estimation based on a posited outcome regression model for the relationship between response and covariates and methods that use fitted models for the propensity score, the probability of the response being observed given covariates (Rosenbaum and Rubin, 1983), such as stratification or matching (Rosenbaum and Rubin, 1984; Rubin and Thomas, 1996; Lunceford and Davidian, 2004) and “inverse probability weighting” of responses (Robins, Rotnitzky, and Zhao, 1994; Rosenbaum,

1987; Lunceford and Davidian, 2004). These methods require correct specification of the model for outcome regression or propensity score, respectively. Robins et al. (1994) identified a class of “augmented inverse probability weighted” estimators that involve modeling both the outcome regression and propensity score, with the efficient member of the class obtained when both models are correct. Scharfstein, Rotnitzky, and Robins (1999) noted that estimators in this class are “doubly robust” in that they are consistent for the true population mean even if one of the outcome regression or propensity score models (but not both) is misspecified. Given the protection afforded by this property, these estimators have been advocated for routine use (Bang and Robins, 2005). However, Kang and Schafer (2007) demonstrated via simulation that the usual doubly robust estimator can be severely biased when both models are misspecified, even if they are “nearly” correct, and that bias is especially problematic when some estimated propensity scores are close to zero, yielding very large “weights.” Estimation based on an outcome regression model only performed much better under misspecification in the Kang-Schafer simulation scenario, leading the authors to warn against use of doubly robust estimators in practice. Tan (2006) discussed alternative approaches to constructing doubly robust estimators that may alleviate some of these difficulties. In this chapter, we propose doubly robust estimators that may yield improved performance relative to existing competitors.

1.2 Existing doubly robust estimators

As in Kang and Schafer (2007), we consider the standard missing data set-up; the spirit of the developments is equally relevant to the causal inference context. Consider n subjects drawn at random from a population of interest, where the ideal, “full” data are (Y_i, X_i) , $i = 1, \dots, n$, independent and identically distributed across i ; Y_i is the response or outcome; and X_i is a vector of covariates. As in Section 1.1, Y_i is not available for all subjects; thus, the data actually observed are independent and identically distributed $(R_i Y_i, R_i, X_i)$, $i = 1, \dots, n$, where $R_i = 1$ or 0 as Y_i is observed or missing. The goal is to estimate the population mean, $\mu = E(Y)$, on the basis of these observed data. Throughout, assume responses are missing at random (Rubin, 1978) in that Y_i and R_i are conditionally independent given X_i .

The propensity score is $P(R = 1 \mid X)$; denote the true propensity score as $\pi_0(X)$. Ordinarily, $\pi_0(X)$ is unknown, and it is customary to posit a (parametric) model; for example, a logistic regression model $\pi(X, \gamma) = \{1 + \exp(\tilde{X}^T \gamma)\}^{-1}$, $\tilde{X} = (1, X^T)^T$. Letting $\hat{\gamma}$ denote the maximum likelihood estimator for γ based on (R_i, X_i) , $i = 1, \dots, n$, it is straightforward to show

(Lunceford and Davidian, 2004) that the “inverse probability weighted” estimators

$$\hat{\mu}_{IPW1} = n^{-1} \sum_{i=1}^n \frac{R_i Y_i}{\pi(X_i, \hat{\gamma})} \quad \text{and} \quad \hat{\mu}_{IPW2} = \left\{ \sum_{i=1}^n \frac{R_i}{\pi(X_i, \hat{\gamma})} \right\}^{-1} \sum_{i=1}^n \frac{R_i Y_i}{\pi(X_i, \hat{\gamma})} \quad (1.1)$$

are consistent for μ if $\pi(X, \gamma)$ is correctly specified; that is, $\pi_0(X) = \pi(X, \gamma_0)$ for some γ_0 . This is because

$$\begin{aligned} E \left\{ \frac{RY}{\pi_0(X)} \right\} &= E \left[E \left\{ \frac{RY}{\pi_0(X)} \mid (Y, X) \right\} \right] = E \left[E \{ R \mid (Y, X) \} \frac{Y}{\pi_0(X)} \right] \\ &= E \left\{ E(R \mid X) \frac{Y}{\pi_0(X)} \right\} = E(Y), \end{aligned}$$

where the third equality follows the missing at random assumption. Note that $\pi(X, \hat{\gamma})$ should be greater than zero almost surely.

Alternatively, because under missing at random $E\{E(Y \mid R = 1, X)\} = E\{E(Y \mid X)\} = E(Y)$, letting $h_0(X)$ denote the true outcome regression $E(Y \mid X)$, it is natural to adopt a model $h(X, \xi)$ for $h_0(X)$, estimate ξ by some $\hat{\xi}$ using the “complete cases” $\{i : R_i = 1\}$, and estimate μ by

$$\hat{\mu}_{OR} = n^{-1} \sum_{i=1}^n h(X_i, \hat{\xi}), \quad (1.2)$$

which is consistent for μ if $h(X, \xi)$ is correctly specified; that is, $h_0(X) = h(X, \xi_0)$ for some ξ_0 , and if $\hat{\xi}$ is consistent for ξ_0 . Because $\hat{\xi}$ is based only on the complete cases, if the distributions of X conditional on $R = 1$ and $R = 0$ differ, (1.2) involves extrapolation.

From Robins et al. (1994) and Tsiatis and Davidian (2007), all estimators for μ that are consistent and asymptotically normal when the propensity score model is correct are asymptotically equivalent to an estimator of the form

$$n^{-1} \sum_{i=1}^n \left\{ \frac{R_i Y_i}{\pi(X_i, \hat{\gamma})} + \frac{R_i - \pi(X_i, \hat{\gamma})}{\pi(X_i, \hat{\gamma})} \mathcal{L}(X_i) \right\} \quad (1.3)$$

for arbitrary $\mathcal{L}(X)$. Estimators in class (1.3) are referred to as “augmented inverse probability weighted” because they have the form of $\hat{\mu}_{IPW1}$ in (1.1) plus an “augmentation” term depending on $\mathcal{L}(X)$; $\hat{\mu}_{IPW1}$ is obtained when $\mathcal{L}(X) \equiv 0$. From Robins et al. (1994), the estimator with the smallest asymptotic variance among those in class (1.3) (so with $\pi(X, \gamma)$ correct) is

$$\hat{\mu}_{DR} = n^{-1} \sum_{i=1}^n \left\{ \frac{R_i Y_i}{\pi(X_i, \hat{\gamma})} - \frac{R_i - \pi(X_i, \hat{\gamma})}{\pi(X_i, \hat{\gamma})} h(X_i, \hat{\xi}) \right\}, \quad (1.4)$$

taking $\mathcal{L}(X_i) = -h(X_i, \hat{\xi})$, where $h(X, \xi)$ is correctly specified, and $\hat{\xi}$ is consistent for ξ_0 . Scharfstein et al. (1999) noted that $\hat{\mu}_{DR}$ remains consistent if only one of the outcome regression model $h(X, \xi)$ or the propensity score model $\pi(X, \gamma)$ is correctly specified, but is inconsistent if both are misspecified; this property is referred to as “double robustness.” If $h(X, \xi)$ is correct, then $\hat{\mu}_{OR}$ is at least as efficient as $\hat{\mu}_{DR}$ (Tan, 2007) but is inconsistent otherwise, while double robustness of (1.4) affords protection against such misspecification.

The estimator (1.4), with γ estimated by maximum likelihood and ξ estimated by ordinary or iteratively reweighted least squares is generally regarded as the “usual” doubly robust estimator. Kang and Schafer (2007) and Tan (2006) identified alternative doubly robust estimators, all involving models for the propensity score and outcome regression and some appearing to have forms outside the “augmented” class (1.3). The former authors attributed poor performance of (1.4) when the propensity or both models are misspecified in part to “inverse weighting” by the propensity score. Tsiatis and Davidian (2007) noted that such alternative estimators can be rewritten in the form (1.3) and used semiparametric theory to argue that poor performance when one or the other model is incorrect may be partly a consequence of the method used to estimate ξ . In Section 1.3, we identify doubly robust estimators from this perspective. When both models are correct and γ is estimated by maximum likelihood, all doubly robust estimators are consistent with the same asymptotic variance; moreover, the asymptotic properties do not depend the method used to estimate ξ (Tan, 2007; Tsiatis and Davidian, 2007).

1.3 Alternative doubly robust estimators

In this section, we focus on estimation of ξ in a posited outcome regression model $h(X, \xi)$, possibly nonlinear in ξ , to identify doubly robust estimators with desirable properties. To fix ideas, we consider first a fully specified propensity score model $\pi(X)$, say, involving no unknown parameters; we relax this shortly. Suppose, for some estimator $\hat{\xi}$ for ξ , we estimate μ by

$$n^{-1} \sum_{i=1}^n \left\{ \frac{R_i Y_i}{\pi(X_i)} - \frac{R_i - \pi(X_i)}{\pi(X_i)} h(X_i, \hat{\xi}) \right\}. \quad (1.5)$$

We now examine how to estimate ξ to achieve the estimator for μ of form (1.5) that is (i) doubly robust and, (ii) if the propensity score is correctly specified, has smallest asymptotic variance among all estimators for μ of form (1.5) using $h(X, \xi)$, even if $h(X, \xi)$ is incorrect.

Suppose first that the propensity score is correct, $\pi(X) = \pi_0(X)$, but $h(X, \xi)$ may or

may not be correctly specified. It is straightforward to show that using any estimator $\hat{\xi}$ in (1.5) leads to a consistent estimator for μ whose asymptotic variance is the same as that of

$$n^{-1} \sum_{i=1}^n \left\{ \frac{R_i Y_i}{\pi_0(X_i)} - \frac{R_i - \pi_0(X_i)}{\pi_0(X_i)} h(X_i, \xi^*) \right\}, \quad (1.6)$$

where ξ^* is the limit in probability of $\hat{\xi}$. Using $\text{var}(\cdot) = E\{\text{var}(\cdot | X, Y)\} + \text{var}\{E(\cdot | X, Y)\}$, this variance is proportional to

$$\text{var} \left\{ \frac{RY}{\pi_0(X)} - \frac{R - \pi_0(X)}{\pi_0(X)} h(X, \xi^*) \right\} = E \left[\frac{1 - \pi_0(X)}{\pi_0(X)} \{Y - h(X, \xi^*)\}^2 \right] + \text{var}(Y). \quad (1.7)$$

A natural objective is to identify the value of ξ^* (and corresponding estimators $\hat{\xi}$ converging in probability to it) such that (1.7) is minimized whether or not $h(X, \xi)$ is correct. Letting $h_\xi(X, \xi) = \partial/\partial\xi\{h(X, \xi)\}$, note that (1.7) is minimized by choosing ξ^* as the solution to

$$E[\{1 - \pi_0(X)\} \pi_0^{-1}(X) \{Y - h(X, \xi^*)\} h_\xi(X, \xi^*)] = 0,$$

or equivalently

$$E \left[\frac{1 - \pi_0(X)}{\pi_0(X)} \{h_0(X) - h(X, \xi^*)\} h_\xi(X, \xi^*) \right] = 0; \quad (1.8)$$

denote this value by ξ_{opt}^* . Note that $\xi_{opt}^* = \xi_0$ when $h(X, \xi)$ is correctly specified.

Consider first the ordinary least squares estimator for ξ , $\hat{\xi}_1$, say, solving

$$n^{-1} \sum_{i=1}^n R_i \{Y_i - h(X_i, \xi)\} h_\xi(X_i, \xi) = 0, \quad (1.9)$$

based on the complete cases. If the propensity score is correct, $\pi(X) = \pi_0(X)$, but $h(X, \xi) \neq h_0(X)$ for any ξ , then the left hand side of (1.9) converges in probability to

$$E[\pi_0(X) \{h_0(X) - h(X, \xi)\} h_\xi(X, \xi)]. \quad (1.10)$$

Then $\hat{\xi}_1$ converges in probability to the value ξ_1 such that (1.10) equals zero; however, comparing (1.10) to (1.8) shows $\xi_1 \neq \xi_{opt}^*$. If the propensity score is incorrect, but the outcome regression model is correct, so that $h(X, \xi_0) = h_0(X)$ for some ξ_0 , then the left hand side of (1.9) again converges to (1.10), and $\xi_1 = \xi_0$, so that $\hat{\xi}_1$ converges in probability to ξ_0 . Thus, the estimator (1.5) for μ using $\hat{\xi}_1$ is doubly robust but does not achieve the minimum variance when the outcome regression model is misspecified. Estimation of ξ by solving (1.9) would most likely be undertaken with continuous Y ; a similar result holds if ξ is estimated via iteratively reweighted least squares, as in the case of a generalized linear model $h(X, \xi)$.

Suppose we consider instead estimating ξ by minimizing the empirical variance of (1.5), $n^{-2} \sum_{i=1}^n [R_i Y_i \pi^{-1}(X_i) - \{R_i - \pi(X_i)\} \pi^{-1}(X_i) h(X_i, \xi)]^2$ in ξ , leading to $\hat{\xi}_2$ solving

$$n^{-1} \sum_{i=1}^n \frac{R_i - \pi(X_i)}{\pi(X_i)} \left\{ \frac{R_i Y_i}{\pi(X_i)} - \frac{R_i - \pi(X_i)}{\pi(X_i)} h(X_i, \xi) \right\} h_\xi(X_i, \xi) = 0. \quad (1.11)$$

If the propensity score is correct but $h(X, \xi) \neq h_0(X)$ for any ξ , then the left hand side of (1.11) converges in probability to an expression of the form (1.8); thus, it follows that $\hat{\xi}_2$ converges in probability to ξ_{opt}^* . When the propensity score is incorrect but the outcome regression model is correct, algebra shows that the left hand side of (1.11) converges to

$$E \left(\left[\frac{\pi_0(X) \{1 - \pi(X)\}}{\pi^2(X)} h_0(X) - \left\{ \frac{\pi_0(X) - 2\pi_0(X)\pi(X) + \pi^2(X)}{\pi^2(X)} h(X, \xi) \right\} \right] h_\xi(X, \xi) \right).$$

The value of ξ setting this equal to zero, to which $\hat{\xi}_2$ converges in probability, is clearly not ξ_0 . Thus, the estimator (1.5) using $\hat{\xi}_2$ achieves minimum variance but is not doubly robust.

These calculations show that using familiar or seemingly intuitive techniques to estimate ξ for use in (1.5) leads to estimators for μ that meet one of conditions (i) or (ii) but not both. To satisfy (i) and (ii) simultaneously, we consider $\hat{\xi}_3$ to be the solution to

$$n^{-1} \sum_{i=1}^n \frac{R_i}{\pi(X_i)} \frac{1 - \pi(X_i)}{\pi(X_i)} \{Y_i - h(X_i, \xi)\} h_\xi(X_i, \xi) = 0, \quad (1.12)$$

which may be viewed as “weighted least squares” based on complete cases with “weights” $\{1 - \pi(X_i)\}/\pi^2(X_i)$. When the propensity score is correct but the outcome regression is not, like that of (1.11), the left hand side of (1.12) converges in probability to an expression of the form (1.8), and hence $\hat{\xi}_3$ converges in probability to ξ_{opt}^* . When the outcome regression is correctly specified and the propensity score is not, the left hand side of (1.12) converges to $E[\pi_0(X) \{1 - \pi(X)\} \pi^{-2}(X) \{h_0(X) - h(X, \xi)\} h_\xi(X, \xi)]$, which equals zero when $\xi = \xi_0$, so that $\hat{\xi}_3$ converges in probability to ξ_0 . Thus, the estimator (1.5) for μ with $\hat{\xi} = \hat{\xi}_3$ is doubly robust and achieves minimum asymptotic variance even if $h(X, \xi)$ is misspecified.

In practice, a parametric propensity score model $\pi(X, \gamma)$ would be posited. Here, we cannot use the above results directly to find an estimator for μ of the form of $\hat{\mu}_{DR}$ in (1.4), where $\hat{\gamma}$ is the maximum likelihood estimator for binary regression, that satisfies conditions (i) and (ii). There is an effect of estimating γ that must be taken into account, so that finding $\hat{\xi}$ converging to the minimizer of (1.7), which assumes $\pi(X)$ is fully specified, does not necessarily lead to minimum

asymptotic variance under a model $\pi(X, \gamma)$ with γ estimated. However, we may exploit the insights gained from the foregoing results, as we now demonstrate.

Let $S_\gamma(R, X, \gamma) = \{R - \pi(X, \gamma)\}[\pi(X, \gamma)\{1 - \pi(X, \gamma)\}]^{-1}\pi_\gamma(X, \gamma)$, be the score for γ , where $\pi_\gamma(X, \gamma) = \partial/\partial\gamma\{\pi(X, \gamma)\}$. From the point of view of semiparametric theory, the class of influence functions (Tsiatis, 2006, Chapter 3) corresponding to estimators for μ of form (1.5), with fully and correctly specified $\pi(X)$ but possibly incorrect $h(X, \xi)$ and using $\hat{\xi}$ converging in probability to some ξ^* , have form $RY/\pi_0(X) - [\{R - \pi_0(X)\}/\pi_0(X)]h(X, \xi^*) - \mu$, whereas those corresponding to estimators of form (1.4) when $\pi(X, \gamma)$ is correctly specified, so that $\pi(X, \gamma_0) = \pi_0(X)$ for some γ_0 , have form

$$\frac{RY}{\pi_0(X)} - \frac{R - \pi_0(X)}{\pi_0(X)}h(X, \xi^*) - \Gamma_0^T(\xi^*)\Sigma_{\gamma\gamma,0}^{-1}S_\gamma(\gamma, R, X, \gamma_0) - \mu \quad (1.13)$$

$$= \frac{RY}{\pi_0(X)} - \frac{R - \pi_0(X)}{\pi_0(X)} \left\{ h(X, \xi^*) + \Gamma_0^T(\xi^*)\Sigma_{\gamma\gamma,0}^{-1} \frac{\pi_{\gamma,0}(X)}{1 - \pi_0(X)} \right\} - \mu, \quad (1.14)$$

where $\pi_{\gamma,0}(X) = \pi_\gamma(X, \gamma_0)$; $\Gamma_0(\xi^*) = E[\pi_{\gamma,0}(X)\{h_0(X) - h(X, \xi^*)\}/\pi_0(X)]$; and $\Sigma_{\gamma\gamma,0} = E(\pi_{\gamma,0}(X)\pi_{\gamma,0}^T(X)/[\pi_0(X)\{1 - \pi_0(X)\}])$, assumed nonsingular. The influence functions (1.13) thus involve an additional term due to estimation of γ , the projection onto the propensity score tangent space, the linear space spanned by the score (Tsiatis, 2006, Theorem 9.1). Because the influence function of an estimator dictates its asymptotic variance, we would like to find $\hat{\xi}$ to substitute in (1.4) converging to ξ_{opt}^{**} , say, that minimizes the variance of (1.13). We do this by considering a class of influence functions containing class (1.13), with elements

$$\begin{aligned} & \frac{RY}{\pi_0(X)} - \frac{R - \pi_0(X)}{\pi_0(X)}h(X, \xi^*) - c^{*T}S_\gamma(\gamma, R, X, \gamma_0) - \mu \\ &= \frac{RY}{\pi_0(X)} - \frac{R - \pi_0(X)}{\pi_0(X)} \left\{ h(X, \xi^*) + c^{*T} \frac{\pi_{\gamma,0}(X)}{1 - \pi_0(X)} \right\} - \mu \end{aligned} \quad (1.15)$$

for arbitrary (ξ^*, c^*) . Identifying the expression in braces in (1.15) as a function of (ξ^*, c^*) with $h(X, \xi^*)$ in (1.7) and (1.8), by analogy to (1.7) and (1.8), $(\xi_{opt}^{**}, c_{opt}^{**})$ solving

$$E \left[\frac{1 - \pi_0(X)}{\pi_0(X)} \left\{ h_0(X) - h(X, \xi^*) - c^{*T} \frac{\pi_{\gamma,0}(X)}{1 - \pi_0(X)} \right\} \begin{pmatrix} h_\xi(X, \xi^*) \\ \frac{\pi_{\gamma,0}(X)}{1 - \pi_0(X)} \end{pmatrix} \right] = 0$$

minimize the variance of (1.15). This yields $c_{opt}^{**} = \Gamma_0^T(\xi_{opt}^{**})\Sigma_{\gamma\gamma,0}^{-1}$, so that (1.15) with $(\xi_{opt}^{**}, c_{opt}^{**})$ substituted has the same form as (1.14), and hence ξ_{opt}^{**} minimizes the variance of (1.13). Thus, an estimator for μ of form (1.4) with the smallest asymptotic variance when $\pi(X, \gamma)$ is correctly

specified but $h(X, \xi)$ may not be is achieved by using $\hat{\xi}$ converging in probability to ξ_{opt}^{**} . By analogy to (1.12), we propose estimating ξ by solving jointly in (ξ, c)

$$\sum_{i=1}^n \left[\frac{R_i}{\pi(X_i, \hat{\gamma})} \frac{1 - \pi(X_i, \hat{\gamma})}{\pi(X_i, \hat{\gamma})} \left(\frac{h_\xi(X, \xi)}{\frac{\pi_\gamma(X_i, \hat{\gamma})}{1 - \pi(X_i, \hat{\gamma})}} \right) \left\{ Y_i - h(X_i, \xi) - c^T \frac{\pi_\gamma(X_i, \hat{\gamma})}{1 - \pi(X_i, \hat{\gamma})} \right\} \right] = 0. \quad (1.16)$$

By an argument entirely similar to that following (1.12), when the propensity model is correct but $h(X, \xi)$ may or may not be, $\hat{\xi}_4$, say, solving (1.16) converges in probability to ξ_{opt}^{**} . When $h(X, \xi)$ is correct but $\pi(X, \gamma)$ is not, assuming that $\hat{\gamma}$ converges in probability to some γ^* , the quantity to which the left hand side of (1.16) converges in probability equals zero when $(\xi, c) = (\xi_0, 0)$. Thus, taking $\hat{\xi} = \hat{\xi}_4$ in (1.4) yields an estimator for μ that is (i) doubly robust and (ii) achieves minimum asymptotic variance when the propensity model is correct. In the sequel, we denote this estimator by $\hat{\mu}_{PROJ}$ and denote the “usual” doubly robust estimator taking $\hat{\xi} = \hat{\xi}_1$, the ordinary least squares estimator for ξ solving (1.9), by $\hat{\mu}_{USUAL}$.

Tan (2006) proposed a doubly robust estimator for μ that is closely related to $\hat{\mu}_{PROJ}$. In the present context, Tan’s estimator is equivalent to modeling $E(Y | X)$ by $h(X, \xi)$ and estimating ξ by ordinary or iteratively reweighted least squares ($\hat{\xi}_1$); replacing $h(X, \xi)$ in (1.4) and (1.16) by $\tilde{h}(X, \tilde{\xi}) = \alpha_0 + \alpha_1 h(X, \xi)$, $\tilde{\xi} = (\alpha_0, \alpha_1, \xi^T)^T$; holding ξ fixed at $\hat{\xi}_1$ and solving (1.16) in (α_0, α_1, c) , where $h_\xi(X, \xi)$ is replaced by $\{1, h(X, \hat{\xi}_1)\}^T$; and substituting the resulting estimates for (α_0, α_1) and $\hat{\xi}_1$ for $\tilde{\xi}$ in (1.4). Denote this estimator by $\hat{\mu}_{TAN}$. If, in constructing $\hat{\mu}_{PROJ}$, we similarly replace $h(X, \xi)$ by $\tilde{h}(X, \tilde{\xi})$ in (1.4) and (1.16), but estimate all elements of $\tilde{\xi}$ simultaneously by solving (1.16) with $h_\xi(X, \xi)$ replaced by $\partial/\partial \tilde{\xi} \{\tilde{h}(X, \tilde{\xi})\}$, then, by the same reasoning as above, the resulting estimator for μ will have asymptotic variance at least as small as that of $\hat{\mu}_{TAN}$ when the propensity score is correct, as this estimator for ξ will converge in probability to the optimal value minimizing this variance, while $\hat{\xi}_1$ used by Tan will not. If $h(X, \xi)$ is correctly specified but $\pi(X, \gamma)$ is not, because the estimator for $\tilde{\xi}$ obtained by either method converges in probability to $(0, 1, \xi_0^T)^T$, both $\hat{\mu}_{TAN}$ and this version of $\hat{\mu}_{PROJ}$ are doubly robust; this would also hold if the true form of $E(Y | X)$ were $\alpha_0 + \alpha_1 h(X, \xi)$ for $(\alpha_0, \alpha_1) \neq (0, 1)$. Thus, although this version of $\hat{\mu}_{PROJ}$ and $\hat{\mu}_{TAN}$ are both doubly robust, the former is at least as efficient as the latter.

All of the estimators $\hat{\mu}_{USUAL}$, $\hat{\mu}_{PROJ}$, and $\hat{\mu}_{TAN}$ involve solving jointly a set of M-estimating equations (Stefanski and Boos, 2002); for example, $\hat{\mu}_{USUAL}$ is found by solving the usual score equation for γ , the ordinary least squares equation (1.9), and the estimating equation

implied by (1.4). Thus, the asymptotic variance of the estimator for μ can be approximated by the usual empirical sandwich technique; see Stefanski and Boos (2002). The resulting estimator for variance will be consistent for the true sampling variance even if one or both of the propensity or outcome regression models is incorrectly specified.

1.4 Enhanced propensity score model

Doubly robust estimators such as $\hat{\mu}_{PROJ}$ that also achieve minimum variance when the propensity model is correct but the outcome regression model may not be should lead to improved performance over $\hat{\mu}_{USUAL}$ under these conditions. However, the problem of “large weights” $1/\pi(X_i, \hat{\gamma})$ can also affect performance; as illustrated by Kang and Schafer (2007), if both models are even mildly misspecified, then $\hat{\mu}_{USUAL}$ may be severely biased due to a few very large “weights.” If the propensity model in particular is slightly misspecified, $\pi(X_i, \hat{\gamma})$ can be erroneously close to zero for some i . We consider an approach to address this issue.

If the propensity score model is correct, we expect that $\sum_{i=1}^n R_i/\pi(X_i, \hat{\gamma}) \approx n$. When the estimated propensities for some observations are close to zero, this quantity can be very different from n . We thus consider propensity models and estimators that impose the restriction that this quantity be equal to n ; if the chosen model is misspecified, this restriction will drive estimated propensities away from zero. We thus propose an “enhanced” propensity score model, given by

$$P(R = 1 \mid X) = \pi(X, \delta, \gamma) = 1 - \frac{\exp(\delta + \tilde{X}^T \gamma)}{1 + \exp(\tilde{X}^T \gamma)}, \quad (1.17)$$

where δ is a scalar parameter. If $\delta = 0$, (1.17) reduces to a usual logistic regression model; otherwise, δ is an “enhancement” imposing the constraint $\sum_{i=1}^n R_i/\pi(X_i, \hat{\delta}, \hat{\gamma}) = n$. This follows because the score for δ is $n - \sum_{i=1}^n R_i/\pi(X_i, \delta, \gamma)$, so that if maximum likelihood is used to estimate $(\delta, \gamma^T)^T$, the constraint is satisfied automatically. Because $\pi(X, \delta, \gamma)$ can take values outside $(0, 1)$, we impose $0 < \pi(X, \delta, \gamma) < 1$ and implement maximum likelihood subject to this restriction, which can be carried out with standard optimization packages.

From a semiparametric theory perspective, it may be shown that use of the “enhanced” model should lead to an increase in efficiency in estimation of μ by any of the methods in Section 1.3 relative to using the logistic regression model with γ alone as long as (1.17) contains $\pi_0(X)$. This follows because the influence functions for these estimators when (1.17) is used involve an additional term relative to those for the same estimators using the model with $\delta = 0$. Those with

the additional term have smaller variance; see Tsiatis (2006, Chapter 9).

1.5 Simulation studies

We carried out several simulation studies to assess performance of the proposed methods under two scenarios. For both scenarios, for each of $n = 200$ and 1000 , we considered the four possible combinations of correct and misspecified outcome regression and propensity score models. For each scenario/setting combination, 1000 Monte Carlo data sets were generated, and the estimators $\hat{\mu}_{OR}$, $\hat{\mu}_{USUAL}$, $\hat{\mu}_{TAN}$, and $\hat{\mu}_{PROJ}$ were calculated for each, where $\hat{\mu}_{PROJ}$ was constructed using $\tilde{m}(X, \tilde{\xi})$ as described in Section 1.3. We also constructed the estimators $\hat{\mu}_{USUAL}^{en}$ and $\hat{\mu}_{PROJ}^{en}$, which are the indicated estimators with the “enhanced” propensity model (1.17) replacing the usual logistic propensity model described below and fitted by constrained maximum likelihood. For each estimator, sandwich standard errors and nominal 95% Wald confidence intervals for μ were calculated. To calculate $\hat{\mu}_{USUAL}^{en}$ and $\hat{\mu}_{PROJ}^{en}$, we used the SAS IML optimizer `nlpqn` (SAS Institute, 2006) to fit the enhanced propensity model.

We duplicated the scenarios in Kang and Schafer (2007) and Tan (2007), which were designed so that, when misspecified, the assumed outcome regression and propensity score models were nonetheless “nearly correct;” our choice of these scenarios allows consideration of the proposed methods in a familiar context that was designed to highlight difference among estimators. Kang and Schafer found that, under their scenario, $\hat{\mu}_{USUAL}$ exhibited severe bias when both models were misspecified but “nearly correct,” while $\hat{\mu}_{OR}$ was not as severely affected, leading the authors to contend that “two wrong models are not necessarily better than one.” Tan modified Kang and Schafer’s scenario slightly and showed that versions of $\hat{\mu}_{TAN}$ offered improvement over $\hat{\mu}_{USUAL}$. For the Kang and Schafer scenario, for each $i = 1, \dots, n$, $Z_i = (Z_{i1}, Z_{i2}, Z_{i3}, Z_{i4})^T$ was generated as standard multivariate normal, and the elements of $X_i = (X_{i1}, X_{i2}, X_{i3}, X_{i4})^T$ were defined as

$$\begin{aligned} X_{i1} &= \exp(Z_{i1}/2), \\ X_{i2} &= Z_{i2}/\{1 + \exp(Z_{i1})\} + 10, \\ X_{i3} &= (Z_{i1}Z_{i3}/25 + 0.6)^3, \\ X_{i4} &= (Z_{i2} + Z_{i4} + 20)^2, \end{aligned}$$

so that Z_i may be expressed in terms of X_i . For each i ,

$$Y_i = 210 + 27.4Z_{i1} + 13.7Z_{i2} + 13.7Z_{i3} + 13.7Z_{i4} + \epsilon_i$$

for ϵ_i standard normal; and R_i was generated as Bernoulli with true propensity

$$\pi_0(X_i) = \text{expit}(-Z_{i1} + 0.5Z_{i2} - 0.25Z_{i3} - 0.1Z_{i4}),$$

where $\text{expit}(u) = e^u / (1 + e^u)$. Correctly specified outcome regression and propensity models were thus achieved when an additive linear regression of Y_i on Z_i and a logistic regression with linear predictor additive in the Z_i for R_i , respectively, were fitted; “nearly” correctly specified models involved fitting these models with X_i replacing Z_i ; see Kang and Schafer (2007). The Tan scenario was identical to that of Kang and Schafer except that $X_{i4} = (Z_{i3} + Z_{i4} + 20)^2$. The true value of the mean is $\mu = 210$.

Results for the Kang and Schafer and Tan scenarios are in Tables 1.1, 1.2, 1.3, and 1.4, respectively. When both models are correct, all estimators perform similarly, and all of the doubly robust estimators show negligible Monte Carlo bias when at least one of the models is correctly specified, as expected. Moreover, $\hat{\mu}_{PROJ}$ and $\hat{\mu}_{PROJ}^{en}$ for the most part exhibit efficiencies no worse or better than those of $\hat{\mu}_{OR}$ and the other doubly robust estimators on the basis of root mean square error and median absolute error, and in particular dominate the others when the outcome regression model is misspecified but the propensity model is correct, consistent with the basis of their construction. When both models are incorrectly specified, $\hat{\mu}_{USUAL}$ shows nonnegligible bias, as observed by Kang and Schafer (2007) and Tan (2007); however, use of the “enhanced” propensity model in $\hat{\mu}_{USUAL}^{en}$ eliminates this behavior. The proposed estimators $\hat{\mu}_{PROJ}$ and $\hat{\mu}_{PROJ}^{en}$ exhibit the best performance in terms of bias and efficiency when both models are misspecified; in Section 1.7.4, we sketch a heuristic argument suggesting that this behavior is not unexpected. Overall, $\hat{\mu}_{PROJ}^{en}$ shows the best performance across the range of settings in both scenarios.

Confidence intervals based on sandwich standard errors based on the doubly robust estimators for the most part attain nominal coverage except when both models are misspecified in the Kang and Schafer scenario; those for $\hat{\mu}_{PROJ}$ and $\hat{\mu}_{PROJ}^{en}$ perform consistently well except in this case. Not unexpectedly, when the outcome regression model is misspecified, confidence intervals based on $\hat{\mu}_{OR}$ can suffer from undercoverage.

1.6 Discussion

Our work complements that of Tan (2006, 2007) and Robins et al. (2007), who also demonstrated that it is possible to identify doubly robust estimators that do not suffer the draw-

backs demonstrated by Kang and Schafer (2007) under model misspecification. We have focused our development on estimation of a single treatment mean in order to demonstrate the approach to developing optimal, doubly robust estimators in Section 1.3 in an accessible context; however, the results are relevant to more complex estimands. In the case where a difference of treatment means is of interest, if one restricts attention to outcome regression models linear in a vector of known functions $g(X)$ for both treatments, then taking the difference of the optimal, doubly robust estimators proposed here will lead to an optimal, doubly robust estimator for the mean difference; see Tan (2006). However, this need not hold in general, e.g., if the posited outcome regression models are nonlinear in their parameters. In this case, it is possible to adapt the approach here to derive directly an optimal, doubly robust estimator for the difference; a sketch of the argument is available Section 1.7.1. The proposed methods may also be adapted to the case of estimation of the parameter in a regression model, where an estimator based on the full data may be derived as the solution to an M-estimating equation; we present such methods in the more general case of monotonely coarsened data in Chapter 2.

Like the stabilized weights discussed by Robins et al. (2000), the enhanced propensity score model proposed in Section 1.4 is an effort to avoid weighting that is too disparate across individuals, leading to instability of the estimator for the mean. In the simple context of estimating a single mean, taking a stabilized weights approach is not possible; accordingly, the proposed enhanced model provides an effective alternative. Other methods, such as truncating or smoothing estimated propensities, may also yield improved performance. More details regarding this can be found in Section 1.7.2.

It is worth noting that, when the outcome regression model is correct but the propensity model is not, attempting to improve efficiency would be fruitless. Here, the optimal estimator is $\hat{\mu}_{OR}$, and the propensity score plays no role; see Tsiatis and Davidian (2007).

Detailed formulae for the asymptotic variance of the estimators in this chapter are available in Section 1.7.3.

1.7 Details

1.7.1 Optimal, doubly robust estimator for a treatment mean difference

It is not necessarily the case that the difference between two optimal estimators for two different treatment means is an optimal estimator for the true difference. Consider the situation where we have observed data (R_i, Y_i, X_i) , $i = 1, \dots, n$, where now R_i is a treatment indicator taking values 0 or 1, and Y_i and X_i are response and covariates, respectively. Note that these observed data are different from those in this chapter, $(R_i Y_i, Y_i, X_i)$. In this setting, we may consider separate estimators for the means for treatments 1 and 0, respectively, as

$$n^{-1} \sum_{i=1}^n \left[\frac{R_i Y_i}{\pi(X_i)} - \{R_i - \pi(X_i)\} l_1(X_i) \right], \quad (1.18)$$

where $l_1(X) \in \mathcal{M}_1$, and

$$n^{-1} \sum_{i=1}^n \left[\frac{(1 - R_i) Y_i}{1 - \pi(X_i)} + \{R_i - \pi(X_i)\} l_0(X_i) \right], \quad (1.19)$$

where $l_0(X) \in \mathcal{M}_0$. Here, \mathcal{M}_0 and \mathcal{M}_1 are spaces of functions depending on X . Alternatively, we may consider an estimator for the difference of means directly, i.e.,

$$n^{-1} \sum_{i=1}^n \left[\frac{R_i Y_i}{\pi(X_i)} - \frac{(1 - R_i) Y_i}{1 - \pi(X_i)} - \{R_i - \pi(X_i)\} l(X_i) \right] \quad (1.20)$$

for $l(X) \in \mathcal{M}$. Consider taking the difference of (1.18) and (1.19) as an estimator of the true mean difference. If \mathcal{M}_1 and \mathcal{M}_0 are linear spaces in known vector-valued functions of X , $g_1(X)$ and $g_0(X)$, say; that is, $l_1(X) = g_1(X)^T \xi_1$ and $l_0(X) = g_0(X)^T \xi_0$, then finding optimal estimators for the means involves minimizing the expectation of the summand squared in each of (1.18) and (1.19). Consider (1.18); this minimization corresponds to finding the projection of $RY/\pi(X)$ onto the linear space whose elements are of the form $\{R - \pi(X)\} l_1(X)$, where $l_1(X) \in \mathcal{M}_1$, and similarly for (1.19). Similarly, if we consider estimating the mean difference using (1.21), finding the optimal such estimator by minimizing the expectation of the summand squared, and \mathcal{M} is a linear space, this is equivalent to projecting $RY/\pi(X) - (1 - R)Y/\{1 - \pi(X)\}$ onto the linear space whose elements are of the form $\{R - \pi(X)\} l(X)$, where $l(X) \in \mathcal{M}$. Here, because projection is a linear operation, the projection of this difference is the difference of the projections of each term separately. Thus, if the spaces \mathcal{M}_1 and \mathcal{M}_0 are the same space and the same as \mathcal{M} , then the difference of the optimal (1.18) and (1.19) will indeed be the same as (1.21). This point is

Table 1.1: Simulation results based on 1000 Monte Carlo replications for the Kang and Schafer scenario. Sample size is 200. Bias is Monte Carlo bias, RMSE is root mean square error, MAE is median of absolute errors, MCSD is Monte Carlo standard deviation, AveSE is average of sandwich standard errors, Cov is Monte Carlo coverage of 95% Wald confidence intervals, OR is outcome regression, and PS is propensity score. Smallest, median, second largest, and largest standard errors for entries in Table 1.1, 1.2 : BIAS,(0.04, 0.08, 0.39, 5.58); AVESE, (0.0008, 0.004, 0.58, 6.64); COV, (0.006, 0.007, 0.015, 0.015).

	Bias	RMSE	MAE	MCSD	AveSE	Cov
$n = 200$						
<i>OR correct, PS correct</i>						
$\hat{\mu}_{OR}$	-0.06	2.51	1.66	2.51	2.56	0.96
$\hat{\mu}_{USUAL}$	-0.06	2.51	1.66	2.51	2.56	0.95
$\hat{\mu}_{PROJ}$	-0.07	2.51	1.69	2.51	2.56	0.95
$\hat{\mu}_{TAN}$	-0.05	2.51	1.68	2.51	2.58	0.95
$\hat{\mu}_{USUAL}^{en}$	-0.06	2.51	1.66	2.51	2.56	0.96
$\hat{\mu}_{PROJ}^{en}$	-0.06	2.51	1.68	2.51	2.58	0.95
<i>OR correct, PS incorrect</i>						
$\hat{\mu}_{OR}$	-0.06	2.51	1.66	2.51	2.56	0.96
$\hat{\mu}_{USUAL}$	-0.05	2.53	1.70	2.53	2.57	0.95
$\hat{\mu}_{PROJ}$	-0.06	2.50	1.68	2.50	2.56	0.96
$\hat{\mu}_{TAN}$	-0.05	2.51	1.67	2.51	2.51	0.96
$\hat{\mu}_{USUAL}^{en}$	-0.06	2.51	1.67	2.51	2.62	0.96
$\hat{\mu}_{PROJ}^{en}$	-0.06	2.51	1.70	2.51	2.63	0.96
<i>OR incorrect, PS correct</i>						
$\hat{\mu}_{OR}$	-0.55	3.29	2.14	3.24	3.24	0.93
$\hat{\mu}_{USUAL}$	0.36	3.53	2.33	3.51	3.22	0.94
$\hat{\mu}_{PROJ}$	-0.06	2.57	1.72	2.57	2.60	0.95
$\hat{\mu}_{TAN}$	0.16	2.88	1.96	2.88	2.81	0.95
$\hat{\mu}_{USUAL}^{en}$	0.54	3.26	2.27	3.22	3.37	0.94
$\hat{\mu}_{PROJ}^{en}$	-0.04	2.57	1.70	2.57	2.85	0.96
<i>OR incorrect, PS incorrect</i>						
$\hat{\mu}_{OR}$	-0.55	3.29	2.14	3.24	3.24	0.93
$\hat{\mu}_{USUAL}$	-5.19	13.26	3.62	12.20	6.54	0.92
$\hat{\mu}_{PROJ}$	-0.39	3.58	2.00	3.55	3.28	0.93
$\hat{\mu}_{TAN}$	-1.77	3.52	2.36	3.05	3.04	0.90
$\hat{\mu}_{USUAL}^{en}$	-1.53	3.51	2.29	3.16	5.48	0.91
$\hat{\mu}_{PROJ}^{en}$	-0.31	3.48	1.89	3.47	3.63	0.94

Table 1.2: Simulation results based on 1000 Monte Carlo replications for the Kang and Schafer scenario. Sample size is 1000. Bias is Monte Carlo bias, RMSE is root mean square error, MAE is median of absolute errors, MCSD is Monte Carlo standard deviation, AveSE is average of sandwich standard errors, Cov is Monte Carlo coverage of 95% Wald confidence intervals, OR is outcome regression, and PS is propensity score. Smallest, median, second largest, and largest standard errors for entries in Table 1.1, 1.2 : BIAS, (0.04, 0.08, 0.39, 5.58); AVESE, (0.0008, 0.004, 0.58, 6.64); COV, (0.006, 0.007, 0.015, 0.015).

	Bias	RMSE	MAE	MCSD	AveSE	Cov
$n = 1000$						
<i>OR correct, PS correct</i>						
$\hat{\mu}_{OR}$	-0.03	1.13	0.73	1.13	1.15	0.95
$\hat{\mu}_{USUAL}$	-0.03	1.13	0.73	1.13	1.15	0.95
$\hat{\mu}_{PROJ}$	-0.03	1.13	0.72	1.13	1.15	0.95
$\hat{\mu}_{TAN}$	-0.03	1.13	0.73	1.12	1.15	0.95
$\hat{\mu}_{USUAL}^{en}$	-0.03	1.13	0.73	1.13	1.15	0.95
$\hat{\mu}_{PROJ}^{en}$	-0.03	1.13	0.72	1.13	1.15	0.95
<i>OR correct, PS incorrect</i>						
$\hat{\mu}_{OR}$	-0.03	1.13	0.73	1.13	1.15	0.95
$\hat{\mu}_{USUAL}$	0.01	1.72	0.74	1.72	1.28	0.95
$\hat{\mu}_{PROJ}$	-0.03	1.13	0.73	1.13	1.15	0.95
$\hat{\mu}_{TAN}$	-0.03	1.13	0.73	1.13	1.15	0.95
$\hat{\mu}_{USUAL}^{en}$	-0.03	1.13	0.73	1.13	1.15	0.95
$\hat{\mu}_{PROJ}^{en}$	-0.03	1.13	0.73	1.13	1.15	0.95
<i>OR incorrect, PS correct</i>						
$\hat{\mu}_{OR}$	-0.78	1.68	1.18	1.49	1.48	0.91
$\hat{\mu}_{USUAL}$	0.12	1.64	1.09	1.64	1.54	0.93
$\hat{\mu}_{PROJ}$	0.01	1.14	0.73	1.14	1.16	0.95
$\hat{\mu}_{TAN}$	0.04	1.27	0.85	1.27	1.26	0.95
$\hat{\mu}_{USUAL}^{en}$	0.24	1.55	1.02	1.53	1.42	0.92
$\hat{\mu}_{PROJ}^{en}$	0.02	1.14	0.73	1.14	1.16	0.95
<i>OR incorrect, PS incorrect</i>						
$\hat{\mu}_{OR}$	-0.78	1.68	1.18	1.49	1.48	0.91
$\hat{\mu}_{USUAL}$	-18.05	177.45	5.25	176.53	16.60	0.61
$\hat{\mu}_{PROJ}$	-1.25	1.78	1.35	1.27	1.24	0.83
$\hat{\mu}_{TAN}$	-1.69	2.24	1.80	1.47	1.43	0.76
$\hat{\mu}_{USUAL}^{en}$	-2.00	2.44	2.08	1.41	1.39	0.69
$\hat{\mu}_{PROJ}^{en}$	-0.96	1.58	1.16	1.25	1.24	0.88

Table 1.3: Simulation results based on 1000 Monte Carlo replications for the Tan scenario. Sample size is 200. Entries are as in Table 1.1. The Tan and Kang and Schafer scenarios are distributionally identical in the “OR correct, PS correct” case. Smallest, median, second largest, and largest standard errors for entries in Table 1.3, 1.4: BIAS, (0.04, 0.08, 0.76, 5.66); AVESE, (0.0008, 0.004, 1.59, 9.78); COV, (0.006, 0.007, 0.010, 0.015).

	Bias	RMSE	MAE	MCSD	AveSE	Cov
$n = 200$						
<i>OR correct, PS correct</i>						
$\hat{\mu}_{OR}$	-0.06	2.51	1.66	2.51	2.56	0.96
$\hat{\mu}_{USUAL}$	-0.06	2.51	1.66	2.51	2.56	0.95
$\hat{\mu}_{PROJ}$	-0.07	2.51	1.69	2.51	2.56	0.95
$\hat{\mu}_{TAN}$	-0.05	2.51	1.68	2.51	2.58	0.95
$\hat{\mu}_{USUAL}^{en}$	-0.06	2.51	1.66	2.51	2.56	0.96
$\hat{\mu}_{PROJ}^{en}$	-0.06	2.51	1.68	2.51	2.58	0.95
<i>OR correct, PS incorrect</i>						
$\hat{\mu}_{OR}$	-0.06	2.51	1.66	2.51	2.56	0.96
$\hat{\mu}_{USUAL}$	-0.04	2.55	1.70	2.55	2.59	0.95
$\hat{\mu}_{PROJ}$	-0.06	2.51	1.69	2.51	2.56	0.95
$\hat{\mu}_{TAN}$	-0.05	2.50	1.65	2.50	2.56	0.96
$\hat{\mu}_{USUAL}^{en}$	-0.06	2.51	1.67	2.51	2.57	0.95
$\hat{\mu}_{PROJ}^{en}$	-0.06	2.51	1.68	2.51	2.62	0.95
<i>OR incorrect, PS correct</i>						
$\hat{\mu}_{OR}$	2.64	4.10	3.02	3.14	3.08	0.88
$\hat{\mu}_{USUAL}$	0.74	3.80	2.44	3.72	3.30	0.93
$\hat{\mu}_{PROJ}$	0.56	2.70	1.76	2.64	2.67	0.95
$\hat{\mu}_{TAN}$	0.64	2.79	1.88	2.72	2.72	0.96
$\hat{\mu}_{USUAL}^{en}$	1.37	3.42	2.33	3.13	3.22	0.91
$\hat{\mu}_{PROJ}^{en}$	0.52	2.69	1.75	2.64	3.14	0.95
<i>OR incorrect, PS incorrect</i>						
$\hat{\mu}_{OR}$	2.64	4.10	3.02	3.14	3.08	0.88
$\hat{\mu}_{USUAL}$	-2.76	24.18	2.76	24.02	7.71	0.95
$\hat{\mu}_{PROJ}$	0.51	2.91	1.90	2.87	2.80	0.95
$\hat{\mu}_{TAN}$	0.94	2.99	1.91	2.84	2.84	0.95
$\hat{\mu}_{USUAL}^{en}$	1.36	3.28	2.18	2.99	3.65	0.93
$\hat{\mu}_{PROJ}^{en}$	0.48	2.86	1.86	2.82	3.02	0.95

Table 1.4: Simulation results based on 1000 Monte Carlo replications for the Tan scenario. Sample size is 1000. Entries are as in Table 1.1. The Tan and Kang and Schafer scenarios are distributionally identical in the “OR correct, PS correct” case. Smallest, median, second largest, and largest standard errors for entries in Table 1.3, 1.4: BIAS, (0.04, 0.08, 0.76, 5.66); AVESE, (0.0008, 0.004, 1.59, 9.78); COV, (0.006, 0.007, 0.010, 0.015).

	Bias	RMSE	MAE	MCS	AveSE	Cov
$n = 1000$						
<i>OR correct, PS correct</i>						
$\hat{\mu}_{OR}$	-0.03	1.13	0.73	1.13	1.15	0.95
$\hat{\mu}_{USUAL}$	-0.03	1.13	0.73	1.13	1.15	0.95
$\hat{\mu}_{PROJ}$	-0.03	1.13	0.72	1.13	1.15	0.95
$\hat{\mu}_{TAN}$	-0.03	1.12	0.73	1.12	1.15	0.95
$\hat{\mu}_{USUAL}^{en}$	-0.03	1.13	0.73	1.13	1.15	0.95
$\hat{\mu}_{PROJ}^{en}$	-0.03	1.13	0.72	1.13	1.15	0.95
<i>OR correct, PS incorrect</i>						
$\hat{\mu}_{OR}$	-0.03	1.13	0.73	1.13	1.15	0.95
$\hat{\mu}_{USUAL}$	0.03	2.07	0.74	2.07	1.32	0.95
$\hat{\mu}_{PROJ}$	-0.03	1.13	0.72	1.13	1.15	0.95
$\hat{\mu}_{TAN}$	-0.03	1.12	0.73	1.12	1.15	0.95
$\hat{\mu}_{USUAL}^{en}$	-0.03	1.13	0.73	1.13	1.15	0.95
$\hat{\mu}_{PROJ}^{en}$	-0.03	1.13	0.73	1.13	1.15	0.95
<i>OR incorrect, PS correct</i>						
$\hat{\mu}_{OR}$	2.31	2.72	2.32	1.43	1.41	0.63
$\hat{\mu}_{USUAL}$	0.18	1.84	1.14	1.84	1.64	0.93
$\hat{\mu}_{PROJ}$	0.19	1.18	0.76	1.16	1.17	0.95
$\hat{\mu}_{TAN}$	0.22	1.23	0.79	1.21	1.21	0.94
$\hat{\mu}_{USUAL}^{en}$	0.53	1.56	1.03	1.47	1.27	0.89
$\hat{\mu}_{PROJ}^{en}$	0.18	1.17	0.74	1.16	1.17	0.94
<i>OR incorrect, PS incorrect</i>						
$\hat{\mu}_{OR}$	2.31	2.72	2.32	1.43	1.41	0.63
$\hat{\mu}_{USUAL}$	-17.89	179.88	2.65	178.98	22.60	0.94
$\hat{\mu}_{PROJ}$	0.13	1.25	0.78	1.25	1.20	0.94
$\hat{\mu}_{TAN}$	0.89	1.61	1.00	1.35	1.31	0.91
$\hat{\mu}_{USUAL}^{en}$	0.73	1.47	0.97	1.28	1.27	0.91
$\hat{\mu}_{PROJ}^{en}$	0.22	1.23	0.78	1.21	1.19	0.94

discussed in Tan (2006). If, however, \mathcal{M}_1 and \mathcal{M}_0 are not the same linear space, then this would not necessarily hold, although one could expand these spaces to comprise the space spanned by both $g_1(X)$ and $g_0(X)$, as discussed by Tan (2006). Likewise, if any of the spaces \mathcal{M}_1 , \mathcal{M}_0 , or \mathcal{M} are nonlinear, that is, e.g., elements of \mathcal{M} are of the arbitrary form $l(X, \xi)$, then again, the difference of optimal (1.18) and (1.19) would not necessarily correspond to the optimal (1.21).

The results in the previous paragraph are relevant whether or not the estimators are doubly robust. We have shown in Section 1.3 that how one estimates ξ determines double robustness. If we wish to find an optimal, doubly robust estimator for the mean difference, the results in this chapter apply directly if the spaces are linear. Specifically, if one takes \mathcal{M}_0 and \mathcal{M}_1 to equal \mathcal{M} , then using the strategy in this chapter to estimate ξ separately for each mean and taking the difference in resulting estimators would lead to the optimal doubly robust estimator for the difference. However, this does not necessarily hold in the nonlinear case. It is possible to construct an optimal doubly robust estimator for the difference of the form (1.21) by adapting the strategy in this chapter to identify an appropriate estimator for ξ . A brief sketch of how one would adapt the strategy for finding the optimal doubly robust estimator for the mean difference of the form (1.21) is given below.

We consider the case where the propensity score is fully specified to demonstrate the idea. Extension to the situation where a parametric propensity score model is postulated and fitted by binary regression techniques would be handled analogously to the developments in Section 1.3.

As in the standard set-up for the usual causal inference problem in an observational point exposure study, we may conceptualize potential outcomes Y_0 and Y_1 corresponding to each treatment, with the observed $Y = RY_1 + (1 - R)Y_0$. The goal is to estimate the difference $\Delta = E(Y_1) - E(Y_0)$ based on the observed data under the usual assumption of no unmeasured confounders, namely, that (Y_0, Y_1) are independent of R given X .

Consider as an estimator for Δ

$$n^{-1} \sum_{i=1}^n \left[\frac{R_i Y_i}{\pi(X_i)} - \frac{(1 - R_i) Y_i}{1 - \pi(X_i)} - \{R_i - \pi(X_i)\} l(X_i, \hat{\xi}) \right] \quad (1.21)$$

for some $l(X, \xi)$ and estimator $\hat{\xi}$ for ξ . Following the ideas in Section 1.3, we wish to identify a value ξ_{opt}^* and an estimator $\hat{\xi}$ converging to it such that the asymptotic variance of (1.21) is minimized when the propensity score is correct and such that (1.21) is doubly robust. Using the assumption of no unmeasured confounders, it is straightforward to deduce that the optimal value ξ_{opt}^* is the

solution in ξ to

$$E \left[\left\{ \frac{Y_1}{\pi_0(X)} + \frac{Y_0}{1 - \pi_0(X)} - l(X, \xi) \right\} \pi_0(X) \{1 - \pi_0(X)\} l_\xi(X, \xi) \right] = 0,$$

or equivalently

$$E \left[\left\{ \frac{l^{(1)}(X)}{\pi_0(X)} + \frac{l^{(0)}(X)}{1 - \pi_0(X)} - l(X, \xi) \right\} \pi_0(X) \{1 - \pi_0(X)\} l_\xi(X, \xi) \right] = 0, \quad (1.22)$$

where $l^{(k)}(X) = E(Y_k | X)$, $k = 0, 1$, are the true potential outcome means conditional on X . Suppose we posit models $l_k(X, \alpha_k)$ for $E(Y_k | X)$, $k = 0, 1$, where the models $l_k(X, \alpha_k)$ are possibly nonlinear in α_k , and take

$$l(X, \xi) = \frac{l_0(X, \alpha_0)}{1 - \pi(X)} + \frac{l_1(X, \alpha_1)}{\pi(X)}, \quad \xi = (\alpha_0^T, \alpha_1^T)^T.$$

Suppose we estimate ξ by solving the estimating equation

$$n^{-1} \sum_{i=1}^n \left[\frac{R_i \{1 - \pi(X_i)\}}{\pi(X_i)} \{Y_i - l_1(X, \alpha_1)\} + \frac{(1 - R_i) \pi(X_i)}{1 - \pi(X_i)} \{Y_i - l_0(X, \alpha_0)\} \right] l_\xi(X, \xi) = 0. \quad (1.23)$$

If the propensity score is correct, $\pi(X) = \pi_0(X)$, but the models $l_k(X, \alpha_k)$, $k = 0, 1$, are not, then it is straightforward to show that the left hand side of (1.23) converges to that of (1.22), so that the estimator $\hat{\xi}$ solving (1.23) converges in probability to ξ_{opt}^* . On the other hand, if the propensity score model is not correct but the models $l_k(X, \alpha_k)$ are in the sense that there are values $\alpha_k^{(0)}$ such that $l_k(X, \alpha_k^{(0)}) = l^{(k)}(X)$, $k = 0, 1$, then the left hand side of (1.23) converges to

$$E \left[\frac{\pi_0(X) \{1 - \pi(X)\}}{\pi(X)} \{l^{(1)}(X) - l_1(X, \alpha_1)\} l_\xi(X, \xi) + \frac{\{1 - \pi_0(X)\} \pi(X)}{1 - \pi(X)} \{l^{(0)}(X) - l_0(X, \alpha_0)\} l_\xi(X, \xi) \right],$$

which equals zero when $\xi = \xi_0 = (\alpha_0^{(0)T}, \alpha_1^{(0)T})^T$, so that $\hat{\xi}$ converges in probability to ξ_0 . This shows that the estimator for Δ in (1.21) is doubly robust and achieves the minimum asymptotic variance even if the models $l_k(X, \alpha_k)$ for $E(Y_k | X)$, $k = 0, 1$, are misspecified.

1.7.2 Enhanced propensity score and stabilized weights

To the best of our knowledge, the stabilized weights of Robins were proposed originally in the context of marginal structural models for the counterfactual mean response to a nondynamic

treatment regime in order to avoid weighting that is too disparate across individuals, leading to instability of the mean estimator. In the marginal structural models setting, one takes the estimating equation that one would solve for the associational relationship between regime and mean response and weights it to account for time-dependent confounding. The associational estimating equation is of the form, using Robins' notation,

$$\sum_{i=1}^n f(\bar{A}_i) \{Y_i - \mu(\bar{A}_i, \xi)\} = 0$$

for some function $f(\bar{A})$. One weights this equation by the product of inverse probabilities that depend on past treatment history \bar{A} and past covariate history \bar{L} . Because $f(\bar{A})$ is arbitrary, one can modify its form in the associational estimating equation above to include extra “stabilization” factors, which by definition depend on \bar{A} only. In contrast, in our problem, because we are only estimating a single mean, there is no analogous $f(\bar{A})$ and hence no analogous “stabilization factor”. Hence, in our simpler context, we also wish to avoid disparate weights and instability but we cannot use the “stability weights” approach. Accordingly, we chose to adopt the restriction that $\sum_{i=1}^n R_i / \pi(X_i, \hat{\gamma}) \approx n$ as an alternative method of achieving less disparate weights; based on our experience, this sum can be very different from n when there are disparate weights. Like the “stabilized weights” discussed by Robins et al. (2000), the enhanced propensity score model proposed in Section 1.4 is an effort to avoid weighting that is too disparate across individuals, leading to instability of the estimator for the mean. In the simple context of estimating a single mean, taking a “stabilized weights” approach is not possible; accordingly, the proposed enhanced model provides an effective alternative. Other methods, such as truncating or smoothing estimated propensities, may also yield improved performance.

1.7.3 Calculation of asymptotic variances

We provide expressions required to calculate the asymptotic variances of the four estimators considered in this chapter: $\hat{\mu}_{OR}$, $\hat{\mu}_{USUAL}$, $\hat{\mu}_{TAN}$ and $\hat{\mu}_{PROJ}$. The asymptotic variance of $\hat{\mu}_{USUAL}^{en}$ and $\hat{\mu}_{PROJ}^{en}$ may be calculated using the same formulæ as for $\hat{\mu}_{USUAL}$ and $\hat{\mu}_{PROJ}$; see below.

The following results are valid for a general propensity score model, including the logistic model or the enhanced model discussed in Section 1.4. In order to streamline presentation of the results, denote the propensity score model by $\pi_i = \pi(X_i, \gamma)$. Similarly, denote the outcome

regression model by $h_i = h(X_i, \xi)$. Let the score corresponding to the propensity score model be $S_{\gamma i} = S_{\gamma}(R_i, X_i, \gamma)$, and define $S_{\gamma \gamma i} = \partial / \partial \gamma^T S_{\gamma}(R_i, X_i, \gamma)$, write $h_{\xi i} = h_{\xi}(X_i, \xi)$, $\pi_{\gamma i} = \pi_{\gamma}(X_i, \gamma)$, $h_{\xi \xi i} = \partial^2 / \partial \xi \partial \xi^T \{h(X_i, \xi)\}$, and $\pi_{\gamma \gamma i} = \partial^2 / \partial \gamma \partial \gamma^T \{\pi(X_i, \gamma)\}$. Let τ be collection of unknown parameters involved in obtaining the estimators for μ ; in particular, $\tau = (\xi^T, \mu)^T$ for $\hat{\mu}_{OR}$, $\tau = (\gamma^T, \xi^T, \mu)^T$ for $\hat{\mu}_{USUAL}$, $\tau = (\gamma^T, \xi^T, c^T, \mu)^T$ for $\hat{\mu}_{PROJ}$, and $\tau = (\gamma^T, \xi^T, \alpha_0, \alpha_1, c^T, \mu)^T$ for $\hat{\mu}_{TAN}$.

The estimator for τ , $\hat{\tau}$, in each case can be obtained by solving a set of M-estimating equations $\sum_{i=1}^n \rho_i(\tau) = 0$ (Stefanski and Boos, 2002), where the last element of $\rho_i(\tau)$ corresponds to the estimating equation for μ . Let $A_n = n^{-1} \sum_{i=1}^n A_i = n^{-1} \sum_{i=1}^n \partial / \partial \tau \{\rho_i(\tau)\}$, and $B_n = n^{-1} \sum_{i=1}^n \rho_i(\tau) \rho_i^T(\tau)$. Following standard theory, the asymptotic covariance matrix of $\hat{\tau}$ can be approximated by the empirical sandwich matrix $V_n = n^{-1} A_n^{-1} B_n (A_n^{-1})^T$. Therefore, the asymptotic variance of the four estimators can be approximated by the last, rightmost diagonal entry of the corresponding matrix V_n . We present the form of $\rho_i(\tau)$ and A_i for each of the estimators, from which the form of V_n may be calculated. The desired diagonal entry of V_n may then be obtained numerically, with the required matrix inversion carried out by standard routines.

For $\hat{\mu}_{OR}$, $\rho_i(\tau)$ is given by

$$\rho_i(\tau) = \begin{pmatrix} h_{\xi i} R_i (Y_i - h_i) \\ h_i - \mu \end{pmatrix},$$

and A_i is given by

$$A_i = \begin{pmatrix} D_{1i} & 0 \\ h_{\xi i}^T & -1 \end{pmatrix},$$

where $D_{1i} = h_{\xi \xi i} R_i (Y_i - h_i) - R_i h_{\xi i} h_{\xi i}^T$.

For $\hat{\mu}_{USUAL}$, $\rho_i(\tau)$ is given by

$$\rho_i(\tau) = \begin{pmatrix} \frac{R_i - \pi_i}{\pi_i (1 - \pi_i)} \pi_{\gamma i} \\ h_{\xi i} R_i (Y_i - h_i) \\ \frac{R_i}{\pi_i} (Y_i - h_i) + h_i - \mu \end{pmatrix},$$

and A_i is given by

$$A_i = \begin{pmatrix} D_{1i} & 0 & 0 \\ 0 & D_{2i} & 0 \\ D_{3i} & D_{4i} & -1 \end{pmatrix},$$

where $D_{1i} = \frac{R_i - \pi_i}{\pi_i(1 - \pi_i)}\pi_{\gamma i} - \frac{\pi_i(1 - \pi_i) + (R_i - \pi_i)(1 - 2\pi_i)}{\{\pi_i(1 - \pi_i)\}^2}\pi_{\gamma i}\pi_{\gamma i}^T$, $D_{2i} = h_{\xi i}R_i(Y_i - h_i) - R_i h_{\xi i} h_{\xi i}^T$, $D_{3i} = -\frac{R_i}{\pi_i^2}(Y_i - h_i)\pi_{\gamma i}^T$, and $D_{4i} = \left(1 - \frac{R_i}{\pi_i}\right)h_{\xi i}^T$.

For $\hat{\mu}_{TAN}$, $\rho_i(\tau)$ is given by

$$\rho_i(\tau) = \begin{pmatrix} \frac{R_i - \pi_i}{\pi_i(1 - \pi_i)}\pi_{\gamma i} \\ h_{\xi i}R_i(Y_i - h_i) \\ \frac{R_i(R_i - \pi_i)}{\pi_i^2}\left(Y_i - \alpha_0 - \alpha_1 h_i - c^T \frac{\pi_{\gamma i}}{1 - \pi_i}\right) \\ \frac{R_i(R_i - \pi_i)}{\pi_i^2}h_i\left(Y_i - \alpha_0 - \alpha_1 h_i - c^T \frac{\pi_{\gamma i}}{1 - \pi_i}\right) \\ \frac{R_i(R_i - \pi_i)}{\pi_i^2}\frac{\pi_{\gamma i}}{1 - \pi_i}\left(Y_i - \alpha_0 - \alpha_1 h_i - c^T \frac{\pi_{\gamma i}}{1 - \pi_i}\right) \\ \frac{R_i}{\pi_i}Y_i - \frac{R_i - \pi_i}{\pi_i}(\alpha_0 + \alpha_1 h_i) - c^T S_{\gamma i} - \mu \end{pmatrix},$$

and the matrix A_i is given by

$$A_i = \begin{pmatrix} D_{1i} & 0 & 0 & 0 & 0 & 0 \\ 0 & D_{2i} & 0 & 0 & 0 & 0 \\ D_{3i} & D_{4i} & D_{5i} & h_i D_{5i} & D_{6i} & 0 \\ h_i D_{3i} & 2h_i D_{4i} & h_i D_{5i} & h_i^2 D_{5i} & h_i D_{6i} & 0 \\ D_{7i} & t D_{4i} & t_i D_{5i} & t_i h_i D_{5i} & t_i D_{6i} & 0 \\ D_{8i} & D_{9i} & D_{10i} & h_i D_{10i} & -S_{\gamma i} & -1 \end{pmatrix},$$

where D_{1i} and D_{2i} are the same as for $\hat{\mu}_{USUAL}$, $D_{3i} = D_{31i} - c^T D_{32i}$,

$$D_{31i} = \frac{-R_i \pi_i^2 - 2R_i(1 - \pi_i)}{\pi_i^3}\pi_{\gamma i}^T\left(Y_i - \alpha_0 - \alpha_1 h_i - c^T \frac{\pi_{\gamma i}}{1 - \pi_i}\right),$$

$$D_{32i} = \frac{1}{(1 - \pi_i)^2}\frac{R_i(1 - \pi_i)}{\pi_i^2}(\pi_{\gamma i}(1 - \pi_i) + \pi_{\gamma i}\pi_{\gamma i}^T),$$

$$D_{4i} = -\alpha_1 \frac{R_i(1 - \pi_i)}{\pi_i^2}h_{\xi i}, D_{5i} = -\frac{R_i(1 - \pi_i)}{\pi_i^2}, D_{6i} = -\frac{R_i(1 - \pi_i)}{\pi_i^2}\frac{\pi_{\gamma i}^T}{1 - \pi_i},$$

$$D_{7i} = \frac{\pi_{\gamma i}}{1 - \pi_i}D_{3i} + D_{32i}, t_i = \frac{\pi_{\gamma i}}{1 - \pi_i}, D_{8i} = -\frac{R_i}{\pi_i^2}\pi_{\gamma i}(Y_i - \alpha_0 - \alpha_1 h_i) - c^T S_{\gamma i},$$

$$D_{9i} = -\alpha_1 \frac{R_i - \pi_i}{\pi_i}h_{\xi i}^T, \text{ and } D_{10i} = -\frac{R_i - \pi_i}{\pi_i}.$$

For $\hat{\mu}_{PROJ}$, $\rho_i(\tau)$ is given by

$$\rho_i(\tau) = \begin{pmatrix} \frac{R_i - \pi_i}{\pi_i(1 - \pi_i)} \pi_{\gamma i} \\ \frac{R_i(R_i - \pi_i)}{\pi_i^2} h_{\xi i} \left(Y_i - h_i - c^T \frac{\pi_{\gamma i}}{1 - \pi_i} \right) \\ \frac{R_i(R_i - \pi_i)}{\pi_i^2} \frac{\pi_{\gamma i}}{1 - \pi_i} \left(Y_i - h_i - c^T \frac{\pi_{\gamma i}}{1 - \pi_i} \right) \\ \frac{R_i}{\pi_i} (Y_i - h_i) + h_i - c^T S_{\gamma i} - \mu \end{pmatrix},$$

and A_i is given by

$$A_i = \begin{pmatrix} D_{1i} & 0 & 0 & 0 \\ \frac{\partial h_i}{\partial \xi} D_{2i} & D_{3i} & D_{4i} & 0 \\ D_{5i} & D_{4i}^T & D_{6i} & 0 \\ D_{7i} & D_{8i} & -S_{\gamma i} & -1 \end{pmatrix},$$

where D_{1i} is the same as that for $\hat{\mu}_{TAN}$, $D_{2i} = D_{21i} - c^T D_{22i}$,

$$\begin{aligned} D_{21i} &= \frac{-R_i \pi_i^2 - 2R_i(1 - \pi_i)}{\pi_i^3} \pi_{\gamma i}^T \left(Y_i - h_i - c^T \frac{\pi_{\gamma i}}{1 - \pi_i} \right), \\ D_{22i} &= \frac{1}{(1 - \pi_i)^2} \frac{R_i(1 - \pi_i)}{\pi_i^2} (\pi_{\gamma i}(1 - \pi_i) + \pi_{\gamma i} \pi_{\gamma i}^T), \\ D_{3i} &= \frac{R_i(1 - \pi_i)}{\pi_i^2} m_{\xi \xi i} \left(Y_i - h_i - c^T \frac{\pi_{\gamma i}}{1 - \pi_i} \right) - \frac{R_i(1 - \pi_i)}{\pi_i^2} h_{\xi i} h_{\xi i}^T, \\ D_{4i} &= -\frac{R_i(1 - \pi_i)}{\pi_i^2} h_{\xi i} \frac{\pi_{\gamma i}^T}{1 - \pi_i}, D_{5i} = \frac{\pi_{\gamma i}}{1 - \pi_i} D_{2i} + D_{22i}, \\ D_{6i} &= -\frac{1}{(1 - \pi_i)^2} \frac{R_i(1 - \pi_i)}{\pi_i^2} \pi_{\gamma i} \pi_{\gamma i}^T, D_{7i} = -\frac{R_i}{\pi_i^2} \pi_{\gamma i}^T (Y_i - h_i) - c^T S_{\gamma \gamma^T i}, \text{ and } D_{8i} = \\ &-\frac{R_i - \pi_i}{\pi_i} h_{\xi i}^T. \end{aligned}$$

In Section 1.3, we also constructed an estimator by replacing $h(X, \xi)$ in $\hat{\mu}_{PROJ}$ by $\tilde{h}(X, \tilde{\xi}) = \alpha_0 + \alpha_1 h(X, \xi)$, $\tilde{\xi} = (\alpha_0, \alpha_1, \xi^T)^T$ and estimating all elements of $\tilde{\xi}$ simultaneously by solving the estimating equations (17). The asymptotic variance of the resulting estimator for μ can be calculated using the same formulæ for $\hat{\mu}_{PROJ}$ with $h(X, \xi)$ and $h_{\xi}(X, \xi)$ in $\rho_i(\tau)$ and A_i replaced by $\tilde{h}(X, \tilde{\xi})$ and $\partial/\partial \tilde{\xi} \{\tilde{h}(X, \tilde{\xi})\}$, respectively.

1.7.4 Asymptotic bias under local misspecification

We argue heuristically that $\hat{\mu}_{PROJ}$ and $\hat{\mu}_{PROJ}^{en}$ may perform well under misspecification of both the propensity and outcome regression models. For the outcome regression model we posit

$E(Y|X) = h(X, \xi)$, which we do not necessarily believe is correct and hence we consider the class of estimators for $\mu = E(Y)$ to be

$$n^{-1} \sum_{i=1}^n \left\{ \frac{R_i Y_i}{\pi_0(X_i)} - \frac{R_i - \pi_0(X_i)}{\pi_0(X_i)} h(X_i, \xi) \right\}. \quad (1.24)$$

Assume we have the correct model for the propensity score $\pi_0(X) = P(R = 1|X)$, we have shown that the optimal choice for ξ is ξ_{opt}^* where ξ_{opt}^* minimizes

$$E \left[\frac{1 - \pi_0(X)}{\pi_0(X)} \{Y - h(X, \xi_{opt}^*)\}^2 \right], \quad (1.25)$$

or equivalently

$$E \left[\frac{1 - \pi_0(X)}{\pi_0(X)} \{Y - h(X, \xi_{opt}^*)\} h_\xi(X, \xi_{opt}^*) \right] = 0. \quad (1.26)$$

Suppose we instead used a misspecified model for $P(R = 1|X)$, say $\pi_n(X) = \pi_0(X) + \theta_n s(X)$, where $\lim_{n \rightarrow \infty} n^{1/2} \theta_n = \tau$. That is we will consider local misspecification.

In that case our estimator would be

$$\hat{\mu}_n = n^{-1} \sum_{i=1}^n \left\{ \frac{R_i Y_i}{\pi_n(X_i)} - \frac{R_i - \pi_n(X_i)}{\pi_n(X_i)} h(X_i, \hat{\xi}_n) \right\}, \quad (1.27)$$

where $\hat{\xi}_n$ would be defined as in this chapter. Because $\pi_n(X)$ converges to $\pi_0(X)$ as $n \rightarrow \infty$, we still would obtain that $\hat{\xi}_n$ still converges in probability to ξ_{opt}^* . Because of contiguity, (1.27) would be asymptotically equivalent to

$$\tilde{\mu}_n = n^{-1} \sum_{i=1}^n \left\{ \frac{R_i Y_i}{\pi_n(X_i)} - \frac{R_i - \pi_n(X_i)}{\pi_n(X_i)} h(X_i, \xi_{opt}^*) \right\}. \quad (1.28)$$

A simple expansion of (1.28) about $\pi_0(X)$ yield that

$$\begin{aligned} \tilde{\mu}_n &= n^{-1} \sum_{i=1}^n \left\{ \frac{R_i Y_i}{\pi_0(X_i)} - \frac{R_i - \pi_0(X_i)}{\pi_0(X_i)} h(X_i, \xi_{opt}^*) \right\} \\ &\quad - n^{-1} \sum_{i=1}^n \frac{\theta_n s(X) R_i}{\pi_0^2(X)} \{Y_i - h(X_i, \xi_{opt}^*)\} + o_p(n^{-1/2}). \end{aligned}$$

Therefore if we consider $n^{1/2}(\hat{\mu}_n - \mu)$ or equivalently $n^{1/2}(\tilde{\mu}_n - \mu)$, then it can be written

as

$$n^{-1/2} \sum_{i=1}^n \left\{ \frac{R_i Y_i}{\pi_0(X_i)} - \frac{R_i - \pi_0(X_i)}{\pi_0(X_i)} h(X_i, \xi_{opt}^*) - \mu \right\} \quad (1.29)$$

$$- n^{1/2} \theta_n n^{-1} \sum_{i=1}^n \frac{R_i}{\pi_0^2(X)} s(X) \{Y_i - h(X_i, \xi_{opt}^*)\} + o_p(1). \quad (1.30)$$

We have already shown that (1.29) converges to a normal distribution with mean zero, and moreover, ξ_{opt}^* was chosen so that the asymptotic variance is minimized among the class of estimators in (1.24). Therefore equation (1.30) denotes the asymptotic bias which by the assumption converges to

$$\begin{aligned} & -\tau E \left[\frac{R}{\pi_0^2(X)} s(X) \{Y - h(X, \xi_{opt}^*)\} \right] \\ & = -\tau E \left[\frac{s(X)}{\pi_0(X)} \{Y - h(X, \xi_{opt}^*)\} \right]. \end{aligned} \quad (1.31)$$

Because of (1.26), we note that

$$E \left[\frac{1 - \pi_0(X)}{\pi_0(X)} \{Y - h(X, \xi_{opt}^*)\} c^T h_\xi(X, \xi_{opt}^*) \right] = 0$$

for any constant c with dimension the same as the dimension of ξ .

It follows that, letting $q_0(X) = [1 - \pi_0(X)] / [\pi_0(X)]^{1/2}$, the asymptotic bias may be written as

$$-\tau E \left\{ \left(s(X) / [\pi_0(X) \{1 - \pi_0(X)\}]^{1/2} - q_0(X) c^T h_\xi(X, \xi_{opt}^*) \right) q_0(X) \{Y - h(X, \xi_{opt}^*)\} \right\},$$

the absolute value of which, by the Cauchy-Schwarz inequality is bounded by

$$\tau \left\{ \inf_c E \left(s(X) / [\pi_0(X) \{1 - \pi_0(X)\}]^{1/2} - q_0(X) c^T h_\xi(X, \xi_{opt}^*) \right)^2 \right\}^{1/2} \quad (1.32)$$

$$\times \left(E[\{q_0(X)\}^2 \{Y - h(X, \xi_{opt}^*)\}^2] \right)^{1/2}. \quad (1.33)$$

Notice that (1.32) is the projection of the element $s(X) / [\pi_0(X) \{1 - \pi_0(X)\}]^{1/2}$ onto the linear space spanned by the vector of elements $q_0(X) h_\xi(X, \xi_{opt}^*)$.

Let us compare these results to the corresponding results had we used the estimators

$$\tilde{\mu}_n^* = n^{-1} \sum_{i=1}^n \left\{ \frac{R_i Y_i}{\pi_n(X_i)} - \frac{R_i - \pi_n(X_i)}{\pi_n(X_i)} h(X_i, \tilde{\xi}_n^*) \right\},$$

where $\tilde{\xi}_n^*$ converges in probability to some $\xi^{**} \neq \xi_{opt}^*$. Using a similar expansion we would obtain that $n^{1/2}(\tilde{\mu}_n^* - \mu)$ is asymptotically equivalent to

$$n^{-1/2} \sum_{i=1}^n \left\{ \frac{R_i Y_i}{\pi_0(X_i)} - \frac{R_i - \pi_0(X_i)}{\pi_0(X_i)} h(X_i, \xi^{**}) - \mu \right\} \quad (1.34)$$

$$-\tau E \left[\frac{s(X)}{\pi_0(X)} \{Y - h(X, \xi^{**})\} \right]. \quad (1.35)$$

We therefore note that firstly the asymptotic variance of (1.34) must be greater than the asymptotic variance of $\tilde{\mu}_n$, and secondly the bias in absolute value is bounded by

$$\tau \left\{ E \left(s(X) / [\pi_0(X) \{1 - \pi_0(X)\}]^{1/2} \right)^2 \right\}^{1/2} \\ \times \left(E[\{q_0(X)\}^2 \{Y - h(X, \xi^{**})\}^2] \right)^{1/2}.$$

The first term in this expression must be at least as large as (1.32), because (1.32) is the projection of $s(X) / [\pi_0(X) \{1 - \pi_0(X)\}]^{1/2}$ onto the linear space spanned by $q_0(X) h_\xi(X, \xi_{opt}^*)$, while the second must be greater than or equal to (1.33) by the definition of ξ_{opt}^* . Thus, the bound on asymptotic bias of $\tilde{\mu}_n^*$ is greater than that for $\tilde{\mu}_n$; moreover, the asymptotic variance of $\tilde{\mu}_n^*$ is of course greater than that of $\tilde{\mu}_n$ by construction.

Although the first result does not guarantee that $\tilde{\mu}_n$ will show smaller bias, it does suggest that smaller bias may obtain in many circumstances, particularly if $s(X) / [\pi_0(X) \{1 - \pi_0(X)\}]^{1/2}$ may be well approximated by a linear combination of $q_0(X) h_\xi(X, \xi_{opt}^*)$.

Chapter 2

Improving Efficiency and Robustness of Doubly Robust Estimators in the Presence of Coarsened Data

2.1 Introduction

Studies in which data are to be collected longitudinally on each participant according to a pre-determined schedule are often complicated by dropout, where some subjects leave the study prematurely and do not return, so that the intended data from the point of dropout onward are missing. Ordinarily, interest focuses on questions that can be addressed within the context of a statistical model describing aspects of the distribution of the full data; i.e., the entire complement of data that would have been collected on each subject had dropout not occurred. It is well understood that failure to take dropout into account in analyses based on the observed data, which are curtailed due to dropout for some participants, can lead to biased inferences on full data model parameters of interest. A vast literature exists on different methods for making valid inferences on these parameters based on the observed data under different assumptions regarding the mechanism governing dropout; see, for example, Hogan, Roy, and Korkontzelou (2004), Philipson, Ho, and Henderson (2008), and Molenberghs and Fitzmaurice (2009) and the extensive references therein.

As a running example, we consider data from AIDS Clinical Trials Group (ACTG) Protocol 175 (Hammer et al., 1996), where subjects infected with human immunodeficiency virus (HIV) were randomized to four antiretroviral regimens with equal probability: zidovudine (ZDV),

ZDV+didanosine, ZDV+zalcitabine, and didanosine. On each, CD4 T-cell count (cells/mm³ blood), a measure of immunologic status, was measured at baseline and, ideally, at 20 ± 5 , 40 ± 5 , 60 ± 5 , and 96 ± 5 weeks post-baseline. A number of baseline covariates were also collected. As the latter three regimens showed no mutual differences, we focus on estimating mean CD4 count at 96 ± 5 weeks for the population of subjects receiving any of the three. Of the 1838 such participants, 12%, 30%, 38%, and 49% had dropped out by each of the visit times, respectively; see Section 2.5. Clearly, the substantial dropout by 96 ± 5 weeks complicates inference on the population mean of interest.

Missingness due to dropout in a longitudinal study is a special case of the general setting of monotonely coarsened data. Coarsening refers to the situation where, for each subject, one of a set of $M + 1$ many-to-one functions of the full data, indexed by $r = 1, \dots, M, \infty$, is observed (Heitjan and Rubin, 1991; Heitjan, 1993; Gill, van der Laan, and Robins, 1997; Tsiatis, 2006). In the case of monotone coarsening, the many-to-one function for any $r = 1, \dots, M$ is itself a many-to-one function of the $(r + 1)$ th function, so that $r = 1$ corresponds to the “most coarsened” data and $r = M$ to the least, and, by convention, ∞ denotes no coarsening at all (the full data are observed). Monotone dropout in a longitudinal study fits into this framework, with r indexing $M + 1$ planned visit times at which data would be collected, where $r = 1$ corresponds to baseline. Here, the coarsened data at level r are the data that would be observed on a subject who is present for the r th visit and then drops out prior to the $(r + 1)$ th visit; see Section 2.2.

Analogous to the notion of missing at random in the case of missing data, the mechanism leading to coarsening is coarsening at random (Heitjan and Rubin, 1991) if, for each r , the probability that, given the full data, the data are coarsened at level r depends only on the coarsened data (so not on data that are not observed at level r). Whether or not the coarsening at random assumption is reasonable in a specific context must of course be critically evaluated by the analyst; when it is plausible, a number of approaches have been proposed for making inference on parameters in a full data model based on the observed, coarsened data. These include likelihood-based methods, where a parametric model for the entire distribution of the full data may be posited, from which the likelihood based on the coarsened data can be deduced without the need to specify the form of the coarsening mechanism (e.g., Birmingham, Rotnitzky, and Fitzmaurice, 2003; Little, 2009). These methods will yield valid inferences as long as the posited full data model is correct, but can lead to bias otherwise. In contrast, inverse probability weighted methods (Robins et al., 1994, 1995; Rotnitzky, Robins, and Scharfstein, 1998; Rotnitzky, 2009) require specification of models for the coarsening probabilities, and the resulting estimators are consistent only if these models are correct

and can be unstable in practice if some probabilities of observing the full data are close to zero. Robins et al. (1994) also identified a class of “augmented” inverse probability weighted estimators that, in the present context, involve (parametric) modeling of both the coarsening probabilities and the conditional expectations of certain functions of the full data given the coarsened data for each level of coarsening; see Section 2.3. The efficient member of this class, that with smallest asymptotic variance, is obtained when both sets of models are correctly specified. Scharfstein et al. (1999) noted that estimators in this class are consistent even if one of the sets of models (but not both) is misspecified. Estimators with this property are referred to as “doubly robust” and have been advocated owing to the protection this feature affords (Bang and Robins, 2005). Bang and Robins (2005) described such a doubly robust estimator in the case of a longitudinal study with dropout and provided simulation evidence demonstrating the doubly robust property; see also Seaman and Copas (2009).

Despite the obvious appeal of the doubly robust property, there has been vigorous criticism of doubly robust estimators. Kang and Schafer (2007) presented simulations in the simple situation of estimation of a population mean from an iid sample with missing response that show that the usual doubly robust estimator can exhibit severe bias when both sets of models are just “slightly” misspecified and/or when some probabilities of observing full data are close to zero and argued against use of doubly robust estimators in practice. In this setting, however, Tan (2006, 2007, 2008) and Chapter 1 showed how to construct doubly robust estimators that do not have these shortcomings. In particular, in Chapter 1 we set out expressly to identify the “best” doubly robust estimator, that with smallest asymptotic variance if the coarsening probabilities are correctly specified regardless of whether or not the conditional expectation models are correct, and demonstrated that these estimators are relatively more efficient and exhibit superior robustness to slight modeling mishaps relative to other doubly robust estimators.

In this chapter, we extend these ideas to the general setting of monotonely coarsened data. In Section 2.2, we introduce notation for the monotone coarsening problem, formalize the coarsening at random assumption, and place the special case of a longitudinal study with missing at random dropout in this context. We formalize the inferential objectives and describe the general form of doubly robust estimators in Section 2.3, and in Section 2.4 propose an improved doubly robust estimator, which we specialize to the setting of a longitudinal study with dropout by demonstration of application of the proposed methods to the ACTG 175 data in Section 2.5. Simulations presented in Section 2.6 exhibit the improved performance of the proposed methods.

2.2 General coarsened data framework and coarsening at random

We follow the presentation in Tsiatis (2006, Section 7.1). Denote the full data by Z ; ideally, then, the data intended to be collected are realizations of Z_1, \dots, Z_n , where $Z_i, i = 1, \dots, n$, are independent and identically distributed (iid). Let C be a discrete coarsening variable taking on the $M + 1$ possible values $1, \dots, M, \infty$ corresponding to $M + 1$ levels of coarsening. When $C = r$, $r = 1, \dots, M$, we only observe $G_r(Z)$, a many-to-one function of Z . When $C = \infty$, we observe $G_\infty(Z) = Z$; i.e., there is no coarsening, and the full data are observed. Under monotone coarsening, $G_r(Z)$ is a many-to-one function of $G_{r+1}(Z)$; i.e., $G_r(Z) = f_r \{G_{r+1}(Z)\}$, $r = 1, \dots, M - 1$. Thus, $G_1(Z)$ are the most coarsened data, $G_2(Z)$ are less so, and so forth, up to $G_\infty(Z) = Z$, where there is no coarsening. The observed data are realizations of iid $\{C_i, G_{C_i}(Z_i)\}$, $i = 1, \dots, n$.

As is customary in general missing data problems, we assume that there is a positive probability of observing the full data, which we express as $P(C = \infty | Z) \geq \epsilon > 0$ almost everywhere. The coarsening at random assumption may be expressed as

$$P(C = r | Z) = \pi\{r, G_r(Z)\}, \quad r = 1, \dots, M, \infty; \quad (2.1)$$

i.e., the probability of coarsening at level r depends on the full data Z only as a function $\pi\{r, G_r(Z)\}$ of the observed data $G_r(Z)$. As $G_\infty(Z) = Z$, write $\pi\{\infty, G_\infty(Z)\} = \pi(\infty, Z)$.

We now demonstrate how data from a longitudinal study with dropout fit into this framework, where we use notation for this setting popularized by Robins and colleagues (e.g., Bang and Robins, 2005). Let L_j be the vector of information collected at visit time t_j , $j = 1, \dots, M + 1$. Let R be a dropout indicator such that, if $R = j$, the subject is last seen at the j th visit, and the observed data are $\bar{L}_j = (L_1, \dots, L_j)$, $j = 1, \dots, M + 1$; write $\bar{L} = \bar{L}_{M+1}$. In the coarsened data framework, the full data Z are thus $Z = G_\infty(Z) = \bar{L}$; the coarsening indicator C corresponds to R , $C = 1, \dots, M, \infty$, where $C = \infty$ is the same as $R = M + 1$; and the coarsened data $G_r(Z) = \bar{L}_r$, $r = 1, \dots, M$. The observed data from a sample of size n are then iid (R_i, \bar{L}_{R_i}) , $i = 1, \dots, n$. Thus, in ACTG 175, $M = 4$, $t_1 = 0$ (baseline); $(t_2, t_3, t_4, t_5) = (20, 40, 60, 96) \pm 5$ weeks; $L_1 = (X, Y_1)$, say, where X are baseline covariates and Y_1 is baseline CD4 count; and $L_j = Y_j$, $j = 2, \dots, M + 1 = 5$, where Y_j is CD4 count collected at t_j . The positivity and coarsening at random (2.1) assumptions become $P(R = M + 1 | \bar{L}) \geq \epsilon > 0$ almost everywhere and $P(R = j | \bar{L}) = \pi(j, \bar{L}_j)$, for $j = 1, \dots, M, M + 1$, respectively.

See Tsiatis (2006) for other special cases of the general coarsened data setting.

2.3 Inferential objective and doubly robust estimators

We suppose that the analyst has specified a semiparametric model for the full data Z corresponding to density $p_Z(z; \beta, \eta)$, say, where β ($p \times 1$) is the finite dimensional parameter of interest; here, then, $p_Z(z; \beta, \eta)$ embodies the features of the full data that the analyst is willing to assume. The goal is to estimate β based on the sample of observed, monotonely coarsened data. Ordinarily, η is an infinite dimensional nuisance parameter representing aspects of the full data distribution about which nothing is assumed. If η were finite dimensional, $p_Z(z; \beta, \eta)$ is a fully parametric model, in which case inference on β based on the observed data under the coarsening at random assumption could be carried out via maximum likelihood techniques. We consider estimators for β calculable under a more general semiparametric model.

For ACTG 175, with $Y = Y_5 = \text{CD4 count at } 96 \pm 5 \text{ weeks}$, $\beta = E(Y)$; with no further assumptions, $p_Z(z; \beta, \eta)$ is completely nonparametric except for the restriction of finite β , and, if the full data were available, the obvious estimator is the sample mean at 96 ± 5 weeks. As a second example, if instead one were willing to assume that $E(Y_j|X) = \beta_0 + \beta_1 t_j + \beta_X^T X$, say, and interest focused on the “slope” β_1 , $\beta = (\beta_0, \beta_1, \beta_X^T)^T$, then $p_Z(z; \beta, \eta)$ would be the semiparametric model that imposes this longitudinal (conditional on X) mean structure, leaving all other aspects of the full data distribution unspecified. Were full data available, β could be estimated by solving a set of generalized estimating equations.

In general, we assume that estimators for β exist based on the full data, defined by $(p \times 1)$ unbiased estimating functions $m(Z, \beta)$; i.e., such that $E\{m(Z, \beta)\} = 0$ for all β (or at least for β in a neighborhood of β_0 , the true value). An estimator would solve $\sum_{i=1}^n m(Z_i, \beta) = 0$ and, under regularity conditions, would be consistent and asymptotically normal by standard M-estimator theory (Stefanski and Boos, 2002). For the sample mean at 96 ± 5 weeks in ACTG 175, $m(Z, \beta) = Y - \beta$; for the slope parameter, $m(Z, \beta)$ would be an estimating function from generalized estimating equations, perhaps involving nuisance parameters in a “working” correlation structure.

We start with the premise that the analyst has fully specified the coarsening probabilities (2.1), a requirement we relax in Section 2.4. For general monotonely coarsened data, the theory of Robins et al. (1994) implies that, under the coarsening at random assumption, if the coarsening probabilities $\pi\{r, G_r(Z)\}$ are in fact correctly specified, members of the class of all regular, asymptotically linear estimators (Tsiatis, 2006, Chapter 3) for β using the observed data solve estimating equations based on an augmented inverse probability weighted estimating function of the

form (Tsiatis, 2006, Chapter 10)

$$\frac{I(C = \infty)m(Z, \beta)}{\pi(\infty, Z)} + \sum_{r=1}^M \frac{dM_c \{r, G_r(Z)\}}{K_r \{G_r(Z)\}} \mathcal{L}_r \{G_r(Z)\}, \quad (2.2)$$

where $\pi \{r, G_r(Z)\}$ is as in (2.1); and the discrete hazard of dropout $\lambda_r \{G_r(Z)\} = P(C = r | C \geq r, Z)$, the cumulative hazard $K_r \{G_r(Z)\} = P(C > r | Z)$, $\mathcal{L}_r \{G_r(Z)\}$ are arbitrary functions of $G_r(Z)$, and $dM_c \{r, G_r(Z)\} = I(C = r) - \lambda_r \{G_r(Z)\} I(C \geq r)$, all for $r = 1, \dots, M$ (Tsiatis, 2006, Theorem 9.2). Under coarsening at random, $K_r \{G_r(Z)\} = 1 - \sum_{j=1}^r \pi \{j, G_j(Z)\}$, $r = 1, \dots, M$; $\lambda_1 \{G_1(Z)\} = \pi \{1, G_1(Z)\}$; and $\lambda_r \{G_r(Z)\} = \pi \{r, G_r(Z)\} / \left[1 - \sum_{j=1}^{r-1} \pi \{j, G_j(Z)\}\right]$, $r = 2, \dots, M$. If $\mathcal{L}_r \{G_r(Z)\} = 0$, $r = 1, \dots, M$, (2.2) reduces to the simple inverse probability weighted estimating function depending only on data from the “complete cases,” subjects for whom full data are observed; accordingly, the second “augmentation” term in (2.2) seeks to improve efficiency of estimation of β by exploiting information from all subjects.

It may be shown that, when the $\pi \{r, G_r(Z)\}$ are correct, estimators for β solving estimating equations based on (2.2) will be consistent and asymptotically normal regardless of the choice of the functions $\mathcal{L}_r \{G_r(Z)\}$, $r = 1, \dots, M$, and that the optimal choice, that leading to the estimator for β in the class with smallest asymptotic variance, is $E \{m(Z, \beta_0) | G_r(Z)\}$; write $m(Z) = m(Z, \beta_0)$ for brevity. As these conditional expectations may not be known, a natural strategy is to model them parametrically via functions $h_r \{G_r(Z), \xi\}$, $r = 1, \dots, M$, where ξ is a finite dimensional parameter. One would then estimate β by solving

$$\sum_{i=1}^n \left[\frac{I(C_i = \infty)m(Z_i, \beta)}{\pi(\infty, Z_i)} + \sum_{r=1}^M \frac{dM_c \{r, G_r(Z_i)\}}{K_r \{G_r(Z_i)\}} h_r \{G_r(Z_i), \hat{\xi}\} \right] = 0, \quad (2.3)$$

where $\hat{\xi}$ is some estimator for ξ . The method of estimating ξ is key; see Section 2.4.

If the coarsening probabilities are correctly specified and $\hat{\xi}$ converges in probability to some ξ^* , say, an estimator for β solving (2.3) will be consistent and asymptotically normal regardless of whether or not $h_r \{G_r(Z), \xi^*\} = E \{m(Z) | G_r(Z)\}$, $r = 1, \dots, M$. Moreover, the form of the asymptotic variance of the estimator will be the same if either $\hat{\xi}$ or ξ^* is used in (2.3) (Tsiatis, 2006, Theorem 10.3), so that the asymptotic variance depends only on ξ^* . If $h_r \{G_r(Z), \xi^*\} = E \{m(Z) | G_r(Z)\}$, $r = 1, \dots, M$ does hold, then the estimator for β will be optimal (have smallest asymptotic variance) among all such estimators. If $h_r \{G_r(Z), \xi^*\} \neq E \{m(Z) | G_r(Z)\}$, $r = 1, \dots, M$, but the coarsening probabilities are misspecified, the estimator for β will still be consistent. Accordingly, such estimators are doubly robust, as only one set of models need be correct to ensure consistency.

In the context of a longitudinal study with dropout, Bang and Robins (2005) described estimators for β that are solutions to (2.3); we present details in the next section.

2.4 Existing and proposed doubly robust estimators

We continue to assume that the coarsening probabilities $\pi\{r, G_r(Z)\}$ are fully specified, which we relax shortly. Different methods for estimating ξ in the models $h_r\{G_r(Z), \xi\}$ will lead to different estimators for β solving (2.3). Bang and Robins (2005) advocate one such method, described later in this section. We seek to define an estimator $\hat{\xi}_{opt}$ for ξ in the spirit of Chapter 1; i.e., that (i) is “optimal” when the $\pi\{r, G_r(Z)\}$, $r = 1, \dots, M, \infty$, are correctly specified, even if the $h_r\{G_r(Z)\}$, $r = 1, \dots, M$, are not, in the sense of yielding an estimator $\hat{\beta}_{opt}$ solving (2.3) with smallest asymptotic variance; and (ii) $\hat{\beta}_{opt}$ is doubly robust. Moreover, $\hat{\xi}_{opt}$ requires no further assumptions beyond specification of the models $h_r\{G_r(Z), \xi\}$.

Denote the true coarsening probabilities as $\pi_0\{r, G_r(Z)\}$, and define the true dropout and cumulative hazards as $\lambda_{r0}\{r, G_r(Z)\}$ and $K_{r0}\{r, G_r(Z)\}$ accordingly, where $K_M\{r, G_r(Z)\} = \pi(\infty, Z)$ and $K_{M0}\{r, G_r(Z)\} = \pi_0(\infty, Z)$; and write $dM_{c0}\{r, G_r(Z_i)\}$ when $\lambda_{r0}\{r, G_r(Z)\}$ is substituted for $\lambda_r\{r, G_r(Z)\}$ in $dM_c\{r, G_r(Z_i)\}$. With the coarsening probabilities correct, whether or not the $h_r\{G_r(Z), \xi\}$ are correct, it is straightforward to deduce that minimizing the variance of an estimator for β solving (2.3) involves minimizing in ξ^*

$$E \left[\frac{I(C = \infty)m(Z)}{\pi_0(\infty, Z)} + \sum_{r=1}^M \frac{dM_{c0}\{r, G_r(Z)\}}{K_{r0}\{r, G_r(Z)\}} h_r\{G_r(Z), \xi^*\} \right]^2, \quad (2.4)$$

where ξ^* is the value to which the estimator $\hat{\xi}$ used converges in probability. Denote this minimizing value by ξ^{opt} . If the models $h_r\{G_r(Z), \xi\}$ are correctly specified, so that there is some ξ_0 such that $h_r\{G_r(Z), \xi_0\} = E\{m(Z)|G_r(Z)\}$, $r = 1, \dots, M$, then in fact $\xi^{opt} = \xi_0$; if not, such a ξ^{opt} still exists. Accordingly, to satisfy (i), we require that the desired $\hat{\xi}_{opt}$ converge in probability to ξ^{opt} . To ensure (ii), when the $h_r\{G_r(Z), \xi\}$ are correctly specified but the coarsening probabilities may not be, $\hat{\xi}_{opt}$ must converge in probability to ξ_0 .

From (2.4), ξ^{opt} must satisfy

$$E \left(\left[\sum_{r=1}^M \frac{dM_{c0}\{G_r(Z), \xi\}}{K_{r0}\{G_r(Z), \xi\}} h_{r\xi}\{G_r(Z), \xi\} \right] \left[\frac{I(C = \infty)m(Z)}{\pi(\infty, Z)} + \sum_{r=1}^M \frac{dM_{c0}\{G_r(Z), \xi\}}{K_{r0}\{G_r(Z), \xi\}} h_r\{G_r(Z), \xi\} \right] \right) = 0,$$

where $h_{r\xi} \{G_r(Z), \xi\}$ is the column vector of partial derivatives of $h_r \{G_r(Z)\}$ with respect to ξ . Using Lemmas 10.1–10.3 of Tsiatis (2006), this expression can be written as

$$E \left[-m(Z) \sum_{r=1}^M \frac{\lambda_{r0} \{G_r(Z), \xi\}}{K_{r0} \{G_r(Z), \xi\}} h_{r\xi} \{G_r(Z), \xi\} + \sum_{r=1}^M \frac{\lambda_{r0} \{G_r(Z), \xi\}}{K_{r0} \{G_r(Z), \xi\}} h_{r\xi} \{G_r(Z), \xi\} h_r \{G_r(Z), \xi\} \right] = 0. \quad (2.5)$$

We now derive an estimator $\hat{\xi}_{opt}$ for ξ that converges to ξ^{opt} satisfying (2.5) when the coarsening probabilities are correctly specified but the models $h_r \{G_r(Z), \xi\}$ may not be and that converges to ξ_0 in the converse situation. We propose estimating ξ by solving estimating equations corresponding to the estimating function

$$\sum_{r=1}^M I(C > r) q_r \{G_r(Z), \xi\} [h_{r+1} \{G_{r+1}(Z), \xi\} - h_r \{G_r(Z), \xi\}], \quad (2.6)$$

where $q_r \{G_r(Z), \xi\}$ is a vector of functions with dimension equal to that of ξ , $I(C > M) = I(C = \infty)$, and $h_{M+1} \{G_{M+1}(Z), \xi\} = m(Z)$; note that (2.6) is a function of the observed data $\{C, G_C(Z)\}$. When $r = M$, the summand in (2.6) is

$$I(C = \infty) q_M \{G_M(Z), \xi\} [m(Z) - h_M \{G_M(Z), \xi\}]. \quad (2.7)$$

First assume that the coarsening probabilities may not be correctly specified, so that the posited probabilities $\pi \{r, G_r(Z)\} \neq \pi_0 \{r, G_r(Z)\}$ for some $r = 1, \dots, M, \infty$, but the models $h_r \{G_r(Z), \xi\}$ are correct. We show that (2.6) is an unbiased estimating function under these conditions; i.e., has mean zero, by a series of iterated conditional expectations. The conditional expectation of (2.7) given Z is

$$K_{M0} \{G_M(Z)\} q_M \{G_M(Z), \xi\} [m(Z) - h_M \{G_M(Z), \xi\}], \quad (2.8)$$

and the conditional expectation of (2.8) given $G_M(Z)$ is

$$K_{M0} \{G_M(Z)\} q_M \{G_M(Z), \xi\} [E \{m(Z) | G_M(Z)\} - h_M \{G_M(Z), \xi\}]. \quad (2.9)$$

Under the correctly specified model $h_M \{G_M(Z), \xi_0\} = E \{m(Z) | G_M(Z)\}$, the expectation of (2.9) is zero when $\xi = \xi_0$, and hence (2.7) has expectation zero at $\xi = \xi_0$. We may similarly argue that an arbitrary summand in (2.6) has expectation zero. The conditional expectation of the r th summand, given Z , is $K_{r0} \{G_r(Z)\} q_r \{G_r(Z), \xi\} [h_{r+1} \{G_{r+1}(Z), \xi\} - h_r \{G_r(Z), \xi\}]$,

which in turn has conditional expectation given $G_r(Z)$ equal to $K_{r0} \{G_r(Z)\} q_r \{G_r(Z), \xi\} \times (E[h_{r+1} \{G_{r+1}(Z), \xi\} | G_r(Z)] - h_r \{G_r(Z), \xi\})$. When the $h_r \{G_r(Z), \xi\}$ are correctly specified, $E[h_{r+1} \{G_{r+1}(Z), \xi_0\} | G_r(Z)] = E[E\{m(Z) | G_{r+1}(Z)\} | G_r(Z)] = E\{m(Z) | G_r(Z)\} = h_r \{G_r(Z), \xi_0\}$, where the second-to-last equality follows by coarsening at random. Hence, at $\xi = \xi_0$, each summand in (2.6) has expectation zero, so that (2.6) is an unbiased estimating function for ξ even if the coarsening probabilities are misspecified, and estimators for ξ based on (2.6) will converge in probability to ξ_0 for arbitrary choice of the functions $q_r \{G_r(Z), \xi\}$. Accordingly, the proposed estimator $\hat{\xi}_{opt}$, which involves a particular choice of these functions, discussed next, converges in probability to ξ_0 under these conditions, as required for (ii).

We now consider the choice of the $q_r \{G_r(Z), \xi\}$ and show that the proposed estimator $\hat{\xi}_{opt}$ using this choice converges in probability to ξ^{opt} when the coarsening probabilities are correctly specified and the functions $h_r \{G_r(Z), \xi\}$ may or may not be. We propose taking

$$q_r \{G_r(Z), \xi\} = -[K_r \{G_r(Z)\}]^{-1} \sum_{j=1}^r \frac{\lambda_j \{G_j(Z)\}}{K_j \{G_j(Z)\}} h_{j\xi} \{G_j(Z), \xi\}, \quad r = 1, \dots, M. \quad (2.10)$$

With (2.10) substituted, we now demonstrate that (2.6) has expectation zero at $\xi = \xi^{opt}$, where ξ^{opt} solves (2.5), when the coarsening probabilities are correctly specified but the functions $h_r \{G_r(Z), \xi\}$ may not be, and hence is an unbiased estimating function under these conditions, so that $\hat{\xi}_{opt}$ converges to ξ^{opt} .

Note that (2.6) may be written as $S_1 + S_2 + S_3$, where

$$\begin{aligned} S_1 &= I(C = \infty) q_M \{G_M(Z), \xi\} m(Z), \\ S_2 &= - \sum_{r=2}^M [I(C > r) q_r \{G_r(Z), \xi\} - I(C > r-1) q_{r-1} \{G_{r-1}(Z), \xi\}] h_r \{G_r(Z), \xi\}, \\ S_3 &= -I(C > 1) q_1 \{G_1(Z), \xi\} h_1 \{G_1(Z), \xi\}. \end{aligned}$$

It is straightforward to show, by substituting (2.10), recalling that $\pi\{r, G_r(Z)\} = \pi_0\{r, G_r(Z)\}$, and first finding $E(S_1|Z)$ using $E\{I(C = \infty)|Z\} = \pi_0(\infty, Z)$, that we have

$$E(S_1) = E[-m(Z) \sum_{j=1}^M \lambda_{j0} \{G_j(Z)\} h_{j\xi} \{G_j(Z)\} / K_{j0} \{G_j(Z)\}],$$

which matches the first term in (2.5). Considering an arbitrary summand S_{2r} , say, in S_2 , and using $E\{I(C > r)|Z\} = K_{r0} \{G_r(Z)\}$, we have

$$E(S_{2r}|Z) = \left[\sum_{j=1}^r \frac{\lambda_{j0} \{G_j(Z)\} h_{j\xi} \{G_j(Z), \xi\}}{K_{j0} \{G_j(Z)\}} - \sum_{j=1}^{r-1} \frac{\lambda_{j0} \{G_j(Z)\} h_{j\xi} \{G_j(Z), \xi\}}{K_{j0} \{G_j(Z)\}} \right] h_r \{G_r(Z), \xi\},$$

so that $E(S_{2r}) = E[\lambda_{r0} \{G_r(Z)\} h_{r\xi} \{G_j(Z), \xi\} / K_{r0} \{G_j(Z)\} h_r \{G_r(Z), \xi\}]$, and thus $E(S_2)$ is equal to the second term in (2.5) except for the summand at $r = 1$. An analogous argument applied to S_3 shows that $E(S_3) = E[\lambda_{10} \{G_1(Z)\} h_{1\xi} \{G_j(Z), \xi\} / K_{10} \{G_j(Z)\} h_1 \{G_r(Z), \xi\}]$. Combining these results, the estimating function found by substituting (2.10) into (2.6) has the same expectation as that of (2.5) when the coarsening probabilities are correctly specified, which is equal to zero when $\xi = \xi^{opt}$. Thus, $\hat{\xi}_{opt}$ converges in probability to ξ^{opt} , ensuring (i).

Summarizing, the proposed estimator $\hat{\beta}_{opt}$, found by using (2.10) in the estimating function (2.6) for ξ to obtain $\hat{\xi}^{opt}$, will be doubly robust and achieve smallest asymptotic variance among estimators in class (2.3) when the coarsening probabilities are correctly specified but the $h_r \{G_r(Z), \xi\}$ may not be. Bang and Robins (2005) proposed an alternative approach, which is effectively equivalent to modeling $E\{m(Z)|G_r(Z)\}$, $r = 1, \dots, M$, by functions $h_r^* \{G_r(Z), \xi_r\}$ corresponding to a generalized linear model with canonical link, where the parameter ξ_r is specific to the r th level of coarsening; and $q_r \{G_r(Z), \xi_r\}$ analogous to those in (2.6) for each r are dictated by the gradient and variance function of the generalized linear model. The resulting estimator for β is doubly robust, but, as it does not exploit the “optimal” choice (2.10) in estimation of the ξ_r , it will not in general achieve the minimum asymptotic variance when the coarsening probabilities are correct unless the $h_r^* \{G_r(Z), \xi_r\}$ are also correct.

The coarsening probabilities are unlikely to be known in practice unless the data arise from a study in which coarsening was by design. Thus, in most applications, it would be natural to postulate and fit parametric models to characterize the coarsening mechanism. As discussed in Tsiatis (2006, Section 8.2), an obvious approach is to model the discrete hazards $\lambda_r \{G_r(Z)\}$, $r = 1, \dots, M$ in terms of a finite dimensional parameter ψ , say, and write $\lambda_r \{G_r(Z), \psi\}$; e.g., via logistic regression; and estimate ψ by maximum likelihood. The likelihood may be shown to be

$$\prod_{r=1}^M \prod_{i: C_i \geq r} [\lambda_r \{G_r(Z_i), \psi\}]^{I(C_i=r)} [1 - \lambda_r \{G_r(Z_i), \psi\}]^{I(C_i > r)},$$

which may be written equivalently as

$$\prod_{i=1}^n \prod_{r=1}^M [\lambda_r \{G_r(Z_i), \psi\}]^{I(C_i=r)} [1 - \lambda_r \{G_r(Z_i), \psi\}]^{I(C_i > r)}$$

(Tsiatis, 2006, Section 8.2). This specification implies corresponding parametric models $\pi\{r, G_r(Z), \psi\}$ and $K_r \{G_r(Z), \psi\}$. Writing $\lambda_{r\psi} \{G_r(Z), \psi\}$ to denote the column vector of partial derivatives of

$\lambda_r \{G_r(Z_i), \psi\}$ with respect to ψ , the score vector for ψ may be shown to be

$$S_\psi \{C, G_C(Z), \psi\} = \sum_{r=1}^M \frac{\lambda_{r\psi} \{G_r(Z), \psi\}}{\lambda_r \{G_r(Z), \psi\} [1 - \lambda_r \{G_r(Z), \psi\}]} dM_C \{r, G_r(Z), \psi\},$$

which, multiplying and dividing by $K_r \{G_r(Z), \psi\}$ and noting that

$$1 - \lambda_r \{G_r(Z), \psi\} = K_r \{G_r(Z), \psi\} / K_{r-1} \{G_r(Z), \psi\},$$

may be written as

$$S_\psi \{C, G_C(Z), \psi\} = \sum_{r=1}^M \frac{dM_C \{r, G_r(Z), \psi\}}{K_r \{G_r(Z), \psi\}} \frac{K_{r-1} \{G_r(Z), \psi\} \lambda_{r\psi} \{G_r(Z), \psi\}}{\lambda_r \{G_r(Z), \psi\}}.$$

As detailed in Tsiatis (2006, Chapters 8-10), there is an effect on the asymptotic distribution of an estimator for β solving (2.3) when the coarsening probabilities are modeled and ψ is estimated by the maximum likelihood estimator $\hat{\psi}$. In particular, it follows from Theorem 9.1 of Tsiatis (2006) that, when the models for the coarsening probabilities are correctly specified, so that there exists ψ_0 such that $\lambda_r \{G_r(Z), \psi_0\} = \lambda_{r0} \{G_r(Z)\}$, the estimator for β that solves the estimating equation

$$\sum_{i=1}^n \left[\frac{I(C_i = \infty) m(Z_i, \beta)}{\pi(\infty, Z_i, \hat{\psi})} + \sum_{r=1}^M \frac{dM_c \{r, G_r(Z_i), \hat{\psi}\}}{K_r \{G_r(Z_i), \hat{\psi}\}} h_r \{G_r(Z_i), \hat{\xi}\} \right] = 0 \quad (2.11)$$

for some $\hat{\xi}$ is asymptotically equivalent to that solving

$$\begin{aligned} & \sum_{i=1}^n \left(\frac{I(C_i = \infty) m(Z_i, \beta)}{\pi(\infty, Z_i, \psi_0)} + \sum_{r=1}^M \left[\frac{dM_c \{r, G_r(Z_i), \psi_0\}}{K_r \{G_r(Z_i), \psi_0\}} h_r \{G_r(Z_i), \xi^*\} - \theta_{proj}^T S_\psi \{C_i, G_C(Z_i), \psi_0\} \right] \right) \\ &= \sum_{i=1}^n \left(\frac{I(C_i = \infty) m(Z_i, \beta)}{\pi(\infty, Z_i, \psi_0)} + \sum_{r=1}^M \frac{dM_c \{r, G_r(Z_i), \psi_0\}}{K_r \{G_r(Z_i), \psi_0\}} \right. \\ & \quad \times \left. \left[h_r \{G_r(Z_i), \xi^*\} - \theta_{proj}^T \frac{K_{r-1} \{G_r(Z_i), \psi_0\} \lambda_{r\psi} \{G_r(Z_i), \psi_0\}}{\lambda_r \{G_r(Z_i), \psi_0\}} \right] \right) = 0, \quad (2.12) \end{aligned}$$

where ξ^* is the limit in probability of $\hat{\xi}$, and θ_{proj} is the value of θ that minimizes

$$E \left[\frac{I(C = \infty) m(Z)}{\pi(\infty, Z, \psi_0)} + \sum_{r=1}^M \frac{dM_c \{r, G_r(Z), \psi_0\}}{K_r \{G_r(Z), \psi_0\}} \tilde{h}_r \{G_r(Z), \tilde{\xi}\} \right]^2, \quad \tilde{\xi} = (\xi^T, \theta^T)^T, \quad (2.13)$$

when ξ^* is substituted for ξ , and

$$\tilde{h}_r \{G_r(Z), \tilde{\xi}\} = h_r \{G_r(Z), \xi\} - \theta^T \frac{K_{r-1} \{G_r(Z), \psi_0\} \lambda_{r\psi} \{G_r(Z), \psi_0\}}{\lambda_r \{G_r(Z), \psi_0\}}. \quad (2.14)$$

Referring back to (2.4), which defines ξ^{opt} assuming ψ_0 is known, and examining (2.12) suggests that the “optimal” estimator for ξ for estimators for β in the class given by (2.11) should converge to the value ξ^{opt*} that minimizes (2.13) simultaneously in ξ^* and θ . In fact, identifying $\tilde{h}_r\{G_r(Z), \tilde{\xi}\}$ in (2.13) with $h_r\{G_r(Z), \xi\}$ in (2.4) shows that the problem of finding the value of ξ^{opt} minimizing (2.4) is analogous to finding the optimal $\tilde{\xi}$, and hence ξ^{opt*} , minimizing (2.13). Accordingly, we can use this correspondence to immediately propose an approach to estimating $\tilde{\xi}$ that will lead to $\hat{\xi}_{opt*}$, say, such that (i) using $\hat{\xi}_{opt*}$ in (2.11) leads to the optimal estimator $\hat{\beta}_{opt*}$ with smallest asymptotic variance among estimators solving (2.11) when the coarsening probabilities are correctly modeled, and (ii) $\hat{\beta}_{opt*}$ is doubly robust.

As, in practice, ψ_0 is unknown, we write $\tilde{h}_r\{G_r(Z), \tilde{\xi}, \psi\}$ to denote (2.14) treating ψ in the coarsening model as a free parameter. With this specification, analogous to (1.15) of Chapter 1, we propose estimating $\tilde{\xi}$ by solving estimating equations corresponding to the estimating function

$$\sum_{r=1}^M I(C > r) \tilde{q}_r \left\{ G_r(Z), \tilde{\xi}, \hat{\psi} \right\} \left[\tilde{h}_{r+1} \left\{ G_{r+1}(Z), \tilde{\xi}, \hat{\psi} \right\} - \tilde{h}_r \left\{ G_r(Z), \tilde{\xi}, \hat{\psi} \right\} \right], \quad (2.15)$$

where $\tilde{q}_r \left\{ G_r(Z), \tilde{\xi}, \psi \right\}$ is the extension of (2.10), namely,

$$\tilde{q}_r \left\{ G_r(Z), \tilde{\xi}, \psi \right\} = -[K_r \left\{ G_r(Z), \psi \right\}]^{-1} \sum_{j=1}^r \frac{\lambda_j \left\{ G_j(Z), \psi \right\}}{K_j \left\{ G_j(Z), \psi \right\}} \begin{bmatrix} \tilde{h}_{j\xi} \left\{ G_j(Z), \tilde{\xi}, \psi \right\} \\ \tilde{h}_{j\theta} \left\{ G_j(Z), \tilde{\xi}, \psi \right\} \end{bmatrix}; \quad (2.16)$$

and

$$\tilde{h}_{j\theta} \left\{ G_j(Z), \tilde{\xi}, \psi \right\} = -K_{j-1} \left\{ G_j(Z), \psi \right\} \lambda_{j\psi} \left\{ G_j(Z), \psi \right\} / \lambda_j \left\{ G_j(Z), \psi \right\}$$

and

$$\tilde{h}_{j\xi} \left\{ G_j(Z), \tilde{\xi}, \psi \right\} = h_{j\xi} \left\{ G_j(Z), \xi, \psi \right\}$$

are column vectors of partial derivatives of (2.14) with respect to θ and ξ .

Noting that $\hat{\psi}$ converges in probability to ψ_0 when the coarsening probabilities are modeled correctly, if they are correct but the $h_r\{G_r(Z), \xi\}$ may not be, by an argument analogous to that following (2.7), we may argue that $\hat{\xi}_{opt*}$ solving the estimating equations defined by (2.15), jointly in θ , will converge to ξ^{opt*} defined above. Likewise, assuming that $\hat{\psi}$ converges to some ψ^* when the coarsening probabilities may not be correct, if this is the case but the models $h_r\{G_r(Z), \xi\}$ are correct, the expectation of (2.15) evaluated at ψ^* may be shown to be equal to zero when $\tilde{\xi} = (\xi, \theta) = (\xi_0, 0)$. Taken together, these results show that (i) and (ii) are satisfied; i.e., the

estimator $\hat{\beta}_{opt*}$ obtained by solving (2.11) with the maximum likelihood estimator $\hat{\psi}$ and the estimator $\hat{\xi}_{opt*}$ solving the estimating equations implied by (2.15) substituted is doubly robust and will have smallest asymptotic variance among all estimators solving (2.11) when the coarsening probabilities are correctly specified but the models $h_r\{G_r(Z), \xi\}$ may not be. Accordingly, $\hat{\beta}_{opt*}$ should be more efficient under the latter conditions than the doubly robust estimator for β of Bang and Robins (2005) obtained when the coarsening probabilities are modeled and ψ is estimated by maximum likelihood, $\hat{\beta}_{br*}$, say.

Because $(\hat{\beta}_{opt*}^T, \hat{\xi}_{opt*}^T, \hat{\psi}^T)^T$ is an M-estimator, and similarly for $\hat{\beta}_{br*}$, the asymptotic covariance matrix for each may be approximated by the empirical sandwich method (Stefanski and Boos, 2002). The resulting estimators will be consistent for the true sampling covariance matrices regardless of whether or not one or both sets of models is misspecified. See Section 2.8.1 for details.

2.5 Application to ACTG 175

We now demonstrate how the foregoing development is specialized to the setting of a longitudinal study with dropout by application to ACTG 175. Referring to the definitions at the end of Section 2.2, recall that interest focuses on estimation of $\beta = E(Y)$, where $Y = Y_5 = \text{CD4 count at } 96 \pm 5 \text{ weeks}$, $M = 4$, and $m(Z, \beta) = Y - \beta$. The baseline covariate vector X includes age (years); weight (kg); Karnofsky score (karnof), an index that reflects ability to perform activities of daily living (scale of 0 to 100); number of days of antiretroviral therapy prior to the trial (antidays); and indicator variables for hemophilia (hemo), homosexual activity (homo), history of intravenous drug use (drug), ZDV within 30 days of the trial, race (0 = white, 1 = nonwhite), gender (0 = female), antiretroviral history (hist; 0 = naive, 1 = experienced), and symptomatic status (symp; 0 = asymptomatic).

We consider estimation of β by three methods: the simple inverse probability weighted estimator, which corresponds to solving (2.11) with all of the $h_r\{G_r(Z), \xi\}$ set equal to zero, denoted $\hat{\beta}_{ipw}$; two versions of the estimator $\hat{\beta}_{br*}$ of Bang and Robins (2005); and two versions of the proposed estimator $\hat{\beta}_{opt*}$. The coarsening at random assumption (2.1) is not unreasonable here; it is widely acknowledged in longitudinal studies of HIV that subjects with certain observed baseline characteristics, such as intravenous drug use, and/or lower evolving CD4 counts prior to dropout, reflecting compromised immunologic status, may be more likely to drop out. Under this assumption, the naive estimator, the sample mean of CD4 counts for the complete cases at 96 ± 5 weeks, is

expected to be biased. The naive sample mean, 348.7 cells/mm³, thus may be an overestimate if, indeed, subjects with poorer disease status are more likely to drop out.

For consistency with the standard notation at the end of Section 2.2, we represent the models we now present by replacing C by R and $G_r(Z)$ by \bar{L}_j and by indexing visits by j in obvious fashion. For use with all estimators, logistic regression models for the discrete hazards at each j were developed. The models involve main effects in elements of \bar{L}_j , which were identified via separate maximum likelihood fits at each j to the data on all subjects with $R \geq j$ using forward selection with entry level of significance 0.15. We also considered other levels of significance, with no qualitative change in results. This yielded models $\lambda_j(\bar{L}_j, \psi) = \text{expit}(\psi_j^T \tilde{\bar{L}}_j)$, $j = 1, \dots, 4$, say, where $\text{expit}(u) = e^u / (1 + e^u)$, $\tilde{\bar{L}}_j$ is the subset of elements of \bar{L}_j selected, and $\psi = (\psi_1^T, \dots, \psi_4^T)^T$. Covariates included in each model are $\tilde{\bar{L}}_1 = (Y_1, \text{age}, \text{drug}, \text{karnof}, \text{antidays}, \text{race}, \text{hist}, \text{symp})$, $\tilde{\bar{L}}_2 = (Y_2, \text{age}, \text{homo}, \text{drug}, \text{antidays}, \text{karnof})$, $\tilde{\bar{L}}_3 = Y_3$, and $\tilde{\bar{L}}_4 = (Y_1, Y_3, \text{hemo}, \text{drug}, \text{karnof}, \text{race})$. Finding the final maximum likelihood estimator $\hat{\psi}$ then reduced to carrying out individual maximum likelihood fits of these models for each j .

Noting that $E\{m(Z)|\bar{L}_j\} = E(Y|\bar{L}_j) - \beta$ for each j , developing models $h_j(\bar{L}_j, \xi)$ and $h_j^*(\bar{L}_j, \xi_j)$, $j = 1, \dots, 4$, for $\hat{\beta}_{opt*}$ and $\hat{\beta}_{br*}$, respectively, corresponds to developing models for the regression of 96±5 week CD4 count on \bar{L}_j ; i.e., for $E(Y|Y_1, \dots, Y_j, X)$. To develop models $h_j(\bar{L}_j, \xi)$, we assumed that the longitudinal data follow the linear mixed model

$$Y_{ij} = \alpha_{0i} + \alpha_{1i}t_{ij} + \gamma^T \tilde{X}_i + e_{ij}, \quad (2.17)$$

where, $\alpha_i = (\alpha_{0i}, \alpha_{1i})^T \sim N\{(\mu_{\alpha 0}, \mu_{\alpha 1})^T, \Sigma_\alpha\}$; $e_{ij} \sim N(0, \sigma_e^2)$ are iid for all i, j ; the α_i , $i = 1, \dots, n$, are independent of each other and all e_{ij} ; and $\tilde{X} = (\text{weight}, \text{karnof}, \text{hist}, \text{symp})$ was identified by fitting (2.17) by maximum likelihood with all of X included and retaining only those elements for which the usual t -test of whether or not the associated coefficient is equal to zero had p-value less than 0.05. Under (2.17), standard results for the multivariate normal distribution yield the required conditional expectations $E(Y|Y_1, \dots, Y_j, X) = E(Y|Y_1, \dots, Y_j, \tilde{X})$, all of which depend on the common $\xi = \{\mu_{\alpha 0}, \mu_{\alpha 1}, \text{vech}(\Sigma_\alpha)^T, \sigma_e^2, \gamma^T\}^T$; see Section 2.8.2. To obtain the first version of $\hat{\beta}_{opt*}$, $\hat{\beta}_{opt*}^{(1)}$, say, we estimated ξ in the implied models $h_j(\bar{L}_j, \xi)$ using (2.15). For direct comparison of the Bang-Robins approach to the proposed method using the same covariate information, we let $h_j^*(\bar{L}_j, \xi_j)$ for each j be linear regression models including main effects in all CD4 counts up through j and \tilde{X} , and estimated the ξ_j by separate ordinary least squares regressions for each j based on the observed data at j ; denote the resulting

estimator by $\hat{\beta}_{br*}^{(1)}$. For a second version of $\hat{\beta}_{br*}$, denoted $\hat{\beta}_{br*}^{(2)}$, we instead considered for each j all of Y_1, \dots, Y_j, X as potential main effects in linear models, and developed and fit these separately by ordinary least squares with forward selection on the elements of X . The resulting $h_j^*(\bar{L}_j, \xi_j)$ contained (age,karnof,race,gender,hist), (age,hemo,drug,karnof,antidays,gender,symp), (age,hemo,karnof,gender), and (age,hemo,karnof) for $j = 1, 2, 3, 4$, respectively, along with (Y_1, \dots, Y_j) . We implemented both $\hat{\beta}_{br*}^{(1)}$ and $\hat{\beta}_{br*}^{(2)}$ as described by Bang and Robins (2005, Section 3). A second version of the proposed estimator, $\hat{\beta}_{opt*}^{(2)}$, was derived by, rather than taking ξ common across j , letting the models implied by (2.17) for each j have j -specific parameters ξ_j . We then let $\xi = (\xi_1^T, \dots, \xi_4^T)^T$, and estimated ξ using (2.15). For all estimators, we obtained standard errors (SEs) via the sandwich technique.

The resulting $\hat{\beta}_{ipw} = 332.96$, (SE 5.10), $\hat{\beta}_{br*}^{(1)} = 333.34$ (4.96), $\hat{\beta}_{opt*}^{(1)} = 333.15$ (4.90), $\hat{\beta}_{br*}^{(2)} = 333.44$ (4.96), and $\hat{\beta}_{opt*}^{(2)} = 333.35$ (4.76). Recognizing that this is a single data set, it is encouraging to note that the estimates are virtually identical, and, consistent with the theory, the inverse probability weighted estimator is inefficient relative to the augmented inverse probability weighted competitors on the basis of estimated SE; moreover, both versions of the proposed estimator achieve or surpass the performance of the Bang and Robins estimators.

We deliberately chose the ACTG 175 study to demonstrate the methods because of a unique feature that highlights explicitly the advantage of consideration of the more general setting of monotone coarsening. Although subjects in the study did cease to attend clinic visits and provide CD4 counts after some time point, so effectively did “drop out” of the study with respect to the response of interest, follow-up of all subjects continued beyond this point. Thus, additional information on each subject throughout the entire 96-week period, regardless of whether or not s/he ceased to attend clinic visits, is available, which we summarize in four time-dependent covariates $\text{dis}_{ij} = I\{\text{subject } i \text{ discontinued study treatment during } (t_j, t_{j+1}]\}$, $j = 1, \dots, 4$; we did not include dis_j in the definitions of L_j in the foregoing analysis for illustrative simplicity, although we certainly could have done so. Acknowledging the availability of these data takes this situation out of the realm of the standard longitudinal dropout setting and notation at the end of Section 2.2, where it is assumed that no data are available beyond visit j if the subject was last seen at j . However, the present setting may still be cast as a case of monotone coarsening and these additional data incorporated in the analysis, as we now demonstrate.

Reverting to the general notation in Section 2.2,

$$Z = (X, Y_1, Y_2, Y_3, Y_4, Y, \text{dis}_1, \text{dis}_2, \text{dis}_3, \text{dis}_4);$$

and, with $C = r$ indicating that the subject last provided a CD4 count at visit r , we observe $G_r(Z) = (X, Y_1, \dots, Y_r, \text{dis}_1, \text{dis}_2, \text{dis}_3, \text{dis}_4)$, $r = 1, \dots, 4$, and $G_\infty(Z) = Z$ for $r = \infty$. Clearly, the coarsened data satisfy the monotonicity requirement. This demonstrates the general principle that one need not think strictly temporally in characterizing monotone missingness in longitudinal data.

We illustrate by calculating $\hat{\beta}_{opt*}$ and $\hat{\beta}_{br*}$ as follows. For both estimators, we derived the discrete hazard models by the same strategy as in the previous analysis, considering all elements of $G_r(Z)$ as possible main effects in the linear predictor of a logistic regression model for each r and retaining a subset by forward selection. This yielded logistic regression models $\lambda_r\{G_r(Z), \psi_r\}$, with main effects for

$$\begin{aligned} &(Y_1, \text{age}, \text{drug}, \text{karnof}, \text{antidays}, \text{race}, \text{hist}, \text{symp}), \\ &(Y_2, \text{age}, \text{homo}, \text{drug}, \text{antidays}, \text{karnof}, \text{dis}_1, \text{dis}_2), \\ &(Y_3, \text{dis}_1, \text{dis}_2), \\ &(Y_1, Y_3, \text{hemo}, \text{karnof}, \text{race}, \text{dis}_2, \text{dis}_4) \end{aligned}$$

for $r = 1, 2, 3, 4$, respectively. To derive models $h_r\{G_r(Z), \xi\}$ for $\hat{\beta}_{opt*}$, we used the form of $E(Y|X, Y_1, \dots, Y_r, \text{dis}_1, \text{dis}_2, \text{dis}_3, \text{dis}_4)$ implied by the linear mixed model $Y_{ir} = \alpha_{0i} + \alpha_{1i}t_{ir} + \gamma^T \tilde{X}_i + \phi_1 I(r \geq 3)\text{dis}_{i2} + \phi_2 I(r = 5)\text{dis}_{i4} + e_{ir}$, where the random effects and within-subject deviations are normal as above, and now $\tilde{X} = (\text{weight}, \text{karnof}, \text{symp})$; see Section 2.8.2. The common $\xi = \{\mu_{\alpha 0}, \mu_{\alpha 1}, \text{vech}(\Sigma_\alpha)^T, \sigma_e^2, \gamma^T, \phi_1, \phi_2\}^T$ was then estimated via (2.15). For $\hat{\beta}_{br*}$, we took $h_r^*\{G_r(Z), \xi_r\} = \gamma_r^T \tilde{X} + \phi_{1,r}\text{dis}_2 + \phi_{2,r}\text{dis}_4 + \zeta_r^T(Y_1, \dots, Y_r)$, so $\xi_r = (\gamma_r^T, \phi_{1,r}, \phi_{2,r}, \zeta_r^T)^T$, which was estimated by ordinary least squares for each r . Using these estimated discrete hazards to also calculate $\hat{\beta}_{ipw}$, $\hat{\beta}_{ipw} = 325.32$ (5.80), $\hat{\beta}_{opt*} = 328.10$ (5.05), and $\hat{\beta}_{br*} = 327.46$ (5.49). As before, performance of the estimators based on estimated SEs is consistent with the theory.

2.6 Simulation studies

We carried out several simulations to assess the performance of the proposed methods in the case of a longitudinal study with dropout, which we describe using the standard notation at the end of Section 2.2. To obtain data for subject i , $i = 1, \dots, n$, we generated baseline covariates corresponding to $t_1 = 0$ as $X_i = (X_{i1}, X_{i2})^T$, where $X_{i1} \sim N(5, 1)$, and $X_{i2} = 0$ or 1 with probability 0.5. For visit times $(t_1, t_2, t_3) = (0, 1, 2)$, we generated longitudinal responses via

the mixed model $Y_{ij} = \alpha_{0i} + \alpha_{1i}t_{ij} + \gamma^T X_i + e_{ij}$, where $(\alpha_{0i}, \alpha_{1i})^T \sim N\{(1.0, 2.5)^T, \Sigma\}$, $\text{vech}(\Sigma) = (0.3, 0.1, 0.2)^T$, $\gamma = (1, -1)^T$, and $e_{ij} \sim N(0, 1)$. Thus, $L_1 = (X, Y_1)$, $L_2 = Y_2$, and $L_3 = Y_3 = Y$. As in ACTG 175, we focus on estimation of $\beta = E(Y) = 10.5$. This set-up implies that, in truth, $E(Y|\bar{L}_1) = \gamma^T X + \mu_1(X, Y_1, Y_2) + t_3\mu_2(X, Y_1, Y_2)$ and $E(Y|\bar{L}_2) = \gamma^T X + \mu_3(X, Y_1) + t_3\mu_4(X, Y_1)$, where the forms of the functions μ_1, \dots, μ_4 are given in Section 2.8.3. Dropout was induced according to the discrete hazards $\lambda_1(\bar{L}_1, \psi) = \text{expit}(\psi_0 + \psi_{1,1}U_1)$ and $\lambda_2(\bar{L}_2, \psi) = \text{expit}(\psi_0 + \psi_{1,2}U_1 + \psi_{2,2}U_2)$, where $U_1 = I(Y_1 > 5.8)$, $U_2 = I(Y_2 > 6.2)$, and $\psi = (\psi_0, \psi_{1,1}, \psi_{1,2}, \psi_{2,2})^T = (-2.0, 2.5, 2.0, 2.5)^T$. This resulted in 36% and 70% missing Y_2 and Y on average, respectively. For each of the following situations, we estimated β by $\hat{\beta}_{ipw}$, $\hat{\beta}_{opt*}$, and $\hat{\beta}_{br*}$.

We considered four scenarios corresponding to each possible combination of specification of correct or incorrect regression models $h_j(\bar{L}_j, \xi)$ and $h_j^*(\bar{L}_j, \xi_j)$ for $E(Y|\bar{L}_j)$, $j = 1, 2$, and correct or incorrect discrete hazard models $\lambda_j(\bar{L}_j, \psi)$. Incorrect discrete hazard models were specified by replacing U_1 and U_2 in the logistic regressions above by Y_1 and Y_2 , respectively. Incorrect models for $E(Y|\bar{L}_j)$ were specified eliminating all terms involving X and replacing Y_1 and Y_2 by $\exp\{(Y_1/9)^2\}$ and $(Y_1 + 3)/\{1 + \exp(Y_2)\} + 1$ in the expressions for μ_1, \dots, μ_4 above. For all methods, the correctly or incorrectly specified discrete hazard models were fit by maximum likelihood. For $\hat{\beta}_{opt*}$, the implied ξ in the correct or incorrect models was estimated based on (2.15); for $\hat{\beta}_{br*}$, the implied ξ_j in these models were estimated by separate ordinary least squares regressions at each j . Each scenario was replicated for $n = 500$ and $n = 1000$, and 1000 Monte Carlo data sets were generated for each sample size-scenario combination.

The results are presented in Table 2.1 and Table 2.2. All estimators show inconsequential bias, although the Monte Carlo biases for $\hat{\beta}_{ipw}$ when the discrete hazards are modeled incorrectly and for $\hat{\beta}_{opt*}$ and $\hat{\beta}_{br*}$ when both sets of models are incorrect for the most part show a trend consistent with the theory. Not surprisingly, when both sets of models are correct, $\hat{\beta}_{opt*}$ and $\hat{\beta}_{br*}$ exhibit virtually identical performance and demonstrated considerable efficiency gains over $\hat{\beta}_{ipw}$. When the discrete hazards are correctly specified but the regression models are not, the proposed estimator shows improved performance on the basis of efficiency over that of Bang and Robins, as expected from its construction. In the converse case, performance of the two estimators is comparable. When both sets of models are incorrect, $\hat{\beta}_{opt*}$ shows a substantial gain in efficiency over $\hat{\beta}_{br*}$ also, consistent with results of Chapter 1 in a much simpler setting. In all cases except when both sets of models are misspecified, leading to estimated SEs that do not reflect the true sampling

Table 2.1: Simulation results based on 1000 Monte Carlo replications. Bias is Monte Carlo bias, RMSE is root mean square error, MCSD is Monte Carlo standard deviation, AveSE is average of sandwich standard errors, Cov is Monte Carlo coverage of 95% Wald confidence intervals, R denotes regression models, and DH denotes discrete hazard models. True value of $\beta = 10.5$. Smallest, median, second largest, and largest standard errors for entries in Table 1 2.1, 2.2: Bias, (0.019, 0.033, 0.067, 0.077); AveSE, (0.004, 0.022, 0.086, 0.340); Cov, (0.007, 0.008, 0.010, 0.011).

	Bias	RMSE	MCSD	AveSE	Cov
$n = 500$					
<i>R correct, DH correct</i>					
$\hat{\beta}_{ipw}$	0.00	1.37	1.37	1.31	0.94
$\hat{\beta}_{br*}$	0.00	0.86	0.86	0.90	0.95
$\hat{\beta}_{opt*}$	-0.01	0.86	0.86	0.89	0.95
<i>R correct, DH incorrect</i>					
$\hat{\beta}_{ipw}$	0.70	2.53	2.43	1.41	0.92
$\hat{\beta}_{br*}$	0.02	1.03	1.03	1.02	0.95
$\hat{\beta}_{opt*}$	0.01	1.04	1.04	1.14	0.94
<i>R incorrect, DH correct</i>					
$\hat{\beta}_{ipw}$	0.00	1.37	1.37	1.31	0.94
$\hat{\beta}_{br*}$	0.04	1.11	1.11	1.01	0.94
$\hat{\beta}_{opt*}$	-0.02	1.04	1.04	1.10	0.94
<i>R incorrect, DH incorrect</i>					
$\hat{\beta}_{ipw}$	0.70	2.53	2.43	1.41	0.92
$\hat{\beta}_{br*}$	0.17	2.52	2.51	1.17	0.94
$\hat{\beta}_{opt*}$	-0.31	1.51	1.48	1.89	0.88

variability, confidence intervals based on all estimators for the most part achieve nominal coverage. Overall, these results indicate that the proposed method performs as advertised and is an attractive alternative to competing estimators.

Table 2.2: Simulation results based on 1000 Monte Carlo replications. Bias is Monte Carlo bias, RMSE is root mean square error, MCSD is Monte Carlo standard deviation, AveSE is average of sandwich standard errors, Cov is Monte Carlo coverage of 95% Wald confidence intervals, R denotes regression models, and DH denotes discrete hazard models. True value of $\beta = 10.5$. Smallest, median, second largest, and largest standard errors for entries in Table 2.1, 2.2: Bias, (0.019, 0.033, 0.067, 0.077); AveSE, (0.004, 0.022, 0.086, 0.340); Cov, (0.007, 0.008, 0.010, 0.011).

	Bias	RMSE	MCSD	AveSE	Cov
$n = 1000$					
<i>R correct, DH correct</i>					
$\hat{\beta}_{ipw}$	-0.03	0.85	0.85	0.87	0.94
$\hat{\beta}_{br*}$	-0.02	0.60	0.60	0.58	0.95
$\hat{\beta}_{opt*}$	-0.01	0.60	0.60	0.61	0.95
<i>R correct, DH incorrect</i>					
$\hat{\beta}_{ipw}$	0.79	2.07	1.91	1.16	0.94
$\hat{\beta}_{br*}$	-0.01	0.61	0.61	0.60	0.95
$\hat{\beta}_{opt*}$	-0.03	0.64	0.64	0.66	0.95
<i>R incorrect, DH correct</i>					
$\hat{\beta}_{ipw}$	-0.03	0.85	0.85	0.87	0.94
$\hat{\beta}_{br*}$	0.01	0.81	0.81	0.83	0.94
$\hat{\beta}_{opt*}$	-0.04	0.72	0.72	0.73	0.94
<i>R incorrect, DH incorrect</i>					
$\hat{\beta}_{ipw}$	0.79	2.07	1.91	1.16	0.94
$\hat{\beta}_{br*}$	0.03	2.11	2.11	1.71	0.90
$\hat{\beta}_{opt*}$	-0.39	1.51	1.46	1.39	0.85

2.7 Discussion

We have proposed an approach to constructing doubly robust estimators for general semi-parametric full data-model parameters of interest based on data that are subject to monotone coarsening under the assumption that the coarsening is at random. A key special case is that of longitudinal data subject to missing at random dropout. The methods extend the ideas of Chapter 1, which focus on estimation of a single population mean response when the response may be missing at random to this more general setting. As in this simpler case, the proposed estimator is designed to equal or exceed the efficiency of other doubly robust estimators, such as those of Bang and Robins

(2005), when models describing the coarsening mechanism are correctly specified, even when regression models that are incorporated to increase efficiency are not. In contrast to dire simulation results presented by Kang and Schafer (2007), our empirical studies show that doubly robust estimators need not exhibit poor performance in this setting, even when both sets of models are incorrectly specified, with our proposed estimator substantially outperforming a competing method on the basis of efficiency, and is no more difficult to implement.

2.8 Details

2.8.1 Derivation of approximate standard errors via the sandwich method

We provide expressions required to calculate the asymptotic variances of the three estimators for β ($p \times 1$) considered in this chapter: $\hat{\beta}_{ipw}$, $\hat{\beta}_{br*}$, and $\hat{\beta}_{opt*}$. Let τ be the collection of unknown parameters involved in obtaining the estimators for β ; in particular, $\tau = (\psi^T, \beta^T)^T$ for $\hat{\beta}_{ipw}$, $\tau = (\psi^T, \xi_1^T, \dots, \xi_M^T, \beta^T)^T$ for $\hat{\beta}_{br*}$, and $\tau = (\psi^T, \xi^T, \theta^T, \beta^T)^T$ for $\hat{\beta}_{opt*}$. The estimator for τ , $\hat{\tau}$, in each case can be obtained by solving a set of M-estimating equations given by $\sum_{i=1}^n \rho_i(\tau) = 0$ (Stefanski and Boos, 2002), where the last p entries of $\rho_i(\tau)$ correspond to the estimating equation for β , and $\rho_i(\tau)$ is defined for each estimator below. Let $A_n = n^{-1} \sum_{i=1}^n A_i = n^{-1} \sum_{i=1}^n \partial/\partial\tau\{\rho_i(\tau)\}$, and $B_n = n^{-1} \sum_{i=1}^n \rho_i(\tau) \rho_i^T(\tau)$. Following standard theory, the asymptotic covariance matrix of $\hat{\tau}$ can be approximated by the empirical sandwich matrix $V_n = n^{-1} A_n^{-1} B_n (A_n^{-1})^T$. Therefore, the asymptotic variances of the three estimators can be approximated by the lower, rightmost diagonal $(p \times p)$ submatrix of the corresponding matrix V_n . We present the form of $\rho_i(\tau)$ and A_i for each of the estimators, from which the form of V_n may be calculated. The desired diagonal submatrix of V_n may then be obtained numerically, with the required matrix inversion carried out by standard routines.

Throughout, we assume that $\lambda_r \{G_r(Z), \psi_r\}$, $r = 1, \dots, M$, are logistic regression models, and $\psi = (\psi_1^T, \dots, \psi_M^T)^T$ are estimated via separate maximum likelihood fits for each $r = 1, \dots, M$, where $\tilde{X}_{i,r}$ is a row vector consisting of the covariates used in the modeling of

$\lambda_r \{G_r(Z_i), \psi_r\}$, including a “1” for the intercept term. For $\hat{\beta}_{ipw}$, $\rho_i(\tau)$ is given by

$$\begin{aligned} \rho_i(\tau) &= \left(\sum_{r=1}^M \frac{dM_C \{r, G_r(Z_i), \psi\}}{K_r \{G_r(Z_i), \psi\}} \frac{K_{r-1} \{G_r(Z_i), \psi\} \lambda_r \psi \{G_r(Z_i), \psi\}}{\lambda_r \{G_r(Z_i), \psi\}} \right) \\ &\quad \frac{I(C_i = \infty) m(Z_i, \beta)}{\pi(\infty, Z_i, \psi)} \\ &= \begin{pmatrix} dM_C \{1, G_1(Z_i), \psi_1\} \tilde{X}_{i,1}^T \\ \vdots \\ dM_C \{M, G_M(Z_i), \psi_M\} \tilde{X}_{i,M}^T \\ \frac{I(C_i = \infty) m(Z_i, \beta)}{\pi(\infty, Z_i, \psi)} \end{pmatrix}, \end{aligned}$$

and A_i is given by

$$A_i = \begin{pmatrix} D_{i,1} & 0 & \cdots & \cdots & \cdots & 0 \\ 0 & \ddots & 0 & \cdots & \cdots & 0 \\ 0 & 0 & D_{i,r} & 0 & \cdots & 0 \\ 0 & \cdots & 0 & \ddots & 0 & 0 \\ 0 & \cdots & \cdots & 0 & D_{i,M} & 0 \\ E_{i,1} & \cdots & E_{i,r} & \cdots & E_{i,M} & D_{i,\beta} \end{pmatrix},$$

where

$$\begin{aligned} D_{i,r} &= -I(C_i \geq r) \lambda_r \{G_r(Z_i), \psi_r\} [1 - \lambda_r \{G_r(Z_i), \psi_r\}] \tilde{X}_{i,r}^T \tilde{X}_{i,r}, \quad r = 1, \dots, M, \\ E_{i,r} &= \frac{I(C_i = \infty) m(Z_i, \beta)}{\pi(\infty, Z_i, \psi)} \lambda_r \{G_r(Z_i), \psi_r\} \tilde{X}_{i,r}, \quad r = 1, \dots, M, \\ D_{i,\beta} &= \frac{I(C_i = \infty)}{\pi(\infty, Z_i, \psi)} m_\beta(Z_i, \beta), \end{aligned}$$

and $m_\beta(Z_i, \beta)$ is a column vector of partial derivatives of $m(Z_i, \beta)$ with respect to β .

We implemented $\hat{\beta}_{br^*}$ as described in Bang and Robins (2005); i.e., we added as a covariate $\hat{K}_r^{-1} \{G_r(Z), \hat{\psi}_1, \dots, \hat{\psi}_r\}$ in the conditional mean functions $h_r^* \{G_r(Z), \xi_r\}$ corresponding to a generalized linear model with canonical link, where $\hat{K}_r^{-1} \{G_r(Z), \hat{\psi}_1, \dots, \hat{\psi}_r\}$ is an estimate for the true cumulative hazard $K_r^{-1} \{G_r(Z), \psi_1, \dots, \psi_r\}$, $r = 1, \dots, M$. We write the new conditional mean function including additional covariate $\hat{K}_r^{-1} \{G_r(Z), \hat{\psi}_1, \dots, \hat{\psi}_r\}$ as

$h_r^* \{G_r(Z), \psi_1, \dots, \psi_r, \xi_r, \beta\}$. For this estimator, $\rho_i(\tau)$ is given by

$$\rho_i(\tau) = \begin{pmatrix} dM_C \{1, G_1(Z_i), \psi_1\} \tilde{X}_{i,1}^T \\ \vdots \\ dM_C \{M, G_M(Z_i), \psi_M\} \tilde{X}_{i,M}^T \\ I(C_i > 1) [h_2^* \{G_2(Z_i), \psi_1, \psi_2, \xi_2, \beta\} - h_1^* \{G_1(Z_i), \psi_1, \xi_1, \beta\}] \\ \quad \times h_{1,\psi_1,\xi_1}^* \{G_1(Z_i), \psi_1, \xi_1, \beta\} \\ \vdots \\ I(C_i > M) [m(Z_i, \beta) - h_M^* \{G_M(Z_i), \psi, \xi_M, \beta\}] \\ \quad \times h_{M,\psi,\xi_M}^* \{G_M(Z_i), \psi, \xi_M, \beta\} \\ h_1^* \{G_1(Z_i), \psi_1, \xi_1, \beta\} \end{pmatrix},$$

where $h_{r,\psi_1,\dots,\psi_r,\xi_r}^* \{G_r(Z_i), \psi_1, \dots, \psi_r, \xi_r, \beta\}$ is the column vector of partial derivatives of $h_r^* \{G_r(Z_i), \psi_1, \dots, \psi_r, \xi_r, \beta\}$ with respect to $\psi_1, \dots, \psi_r, \xi_r$, $r = 1, \dots, M$.

The matrix A_i is given by

$$A_i = \begin{pmatrix} A_{1i} \\ A_{2i} \\ A_{3i} \end{pmatrix}, \quad A_{1i} = \begin{pmatrix} D_{i,1} & 0 & \dots & \dots & \dots & 0 \\ 0 & \ddots & 0 & \dots & \dots & 0 \\ 0 & 0 & D_{i,r} & 0 & \dots & 0 \\ 0 & \dots & 0 & \ddots & 0 & 0 \\ 0 & \dots & \dots & 0 & D_{i,M} & 0 \end{pmatrix},$$

$$\text{and } A_{3i} = \begin{pmatrix} F_{1,i} & 0 & \dots & 0 & F_{2,i} & 0 & \dots & 0 & F_{3,i} \end{pmatrix},$$

where

$$\begin{aligned} D_{i,r} &= -I(C_i \geq r) \lambda_r \{G_r(Z_i), \psi_r\} [1 - \lambda_r \{G_r(Z_i), \psi_r\}] \tilde{X}_{i,r}^T \tilde{X}_{i,r}, \quad r = 1, \dots, M, \\ F_{1,i} &= h_{1,\psi_1}^* \{G_1(Z_i), \psi_1, \xi_1, \beta\}, \quad F_{2,i} = h_{1,\xi_1}^* \{G_1(Z_i), \psi_1, \xi_1, \beta\}, \\ F_{3,i} &= h_{1,\beta}^* \{G_1(Z_i), \psi_1, \xi_1, \beta\}; \end{aligned}$$

i.e., $F_{1,i}, F_{2,i}, F_{3,i}$ are partial derivatives of $h_1^* \{G_1(Z_i), \psi_1, \xi_1, \beta\}$ with respect to ψ_1, ξ_1 , and β ,

respectively. The A_{2i} term involves the partial derivatives of the column vector

$$\rho_{2,i}(\tau) = \begin{pmatrix} I(C_i > 1) [h_2^* \{G_2(Z_i), \psi_1, \psi_2, \xi_2, \beta\} - h_1^* \{G_1(Z_i), \psi_1, \xi_1, \beta\}] \\ \times h_{1,\psi_1,\xi_1}^* \{G_1(Z_i), \psi_1, \xi_1, \beta\} \\ \vdots \\ I(C_i > M) [m(Z_i, \beta) - h_M^* \{G_M(Z_i), \psi, \xi_M, \beta\}] \\ \times h_{M,\psi,\xi_M}^* \{G_M(Z_i), \psi, \xi_M, \beta\} \end{pmatrix}$$

with respect to τ . Often in practice, it is cumbersome to obtain the analytical derivatives of $\rho_{2,i}(\tau)$ with respect to τ . In our implementation, we used numerical derivatives as an approximation to the analytical derivatives. For example, to calculate the derivative of $\rho_{2,i}(\tau)$ with respect to the k th element of τ , we used a one-sided numerical approximation of the form $\{\rho_{2,i}(\tau + \epsilon 1_k) - \rho_{2,i}(\tau)\} / \epsilon$ for small enough $\epsilon > 0$, where 1_k is a column vector with 1 on the k th entry and all other entries 0.

For $\hat{\beta}_{opt^*}$, $\rho_i(\tau)$ is given by

$$\rho_i(\tau) = \begin{pmatrix} dM_C \{1, G_1(Z_i), \psi_1\} \tilde{X}_{i,1}^T \\ \vdots \\ dM_C \{M, G_M(Z_i), \psi_M\} \tilde{X}_{i,M}^T \\ \sum_{r=1}^M I(C_i > r) \tilde{q}_r \{G_r(Z_i), \tilde{\xi}, \psi\} [\tilde{h}_{r+1} \{G_{r+1}(Z_i), \tilde{\xi}, \psi\} - \tilde{h}_r \{G_r(Z_i), \tilde{\xi}, \psi\}] \\ \frac{I(C_i = \infty) m(Z_i, \beta)}{\pi(\infty, Z_i, \psi)} + \sum_{r=1}^M \frac{dM_c \{r, G_r(Z_i), \psi\}}{K_r \{G_r(Z_i), \psi\}} h_r \{G_r(Z_i), \xi\} \end{pmatrix},$$

where

$$\begin{aligned} \tilde{h}_r \{G_r(Z_i), \tilde{\xi}\} &= h_r \{G_r(Z_i), \xi\} - \theta^T \frac{K_{r-1} \{G_r(Z_i), \psi\} \lambda_{r\psi} \{G_r(Z_i), \psi\}}{\lambda_r \{G_r(Z_i), \psi\}}, \\ \tilde{q}_r \{G_r(Z_i), \tilde{\xi}, \psi\} &= -[K_r \{G_r(Z_i), \psi\}]^{-1} \sum_{j=1}^r \frac{\lambda_j \{G_j(Z_i), \psi\}}{K_j \{G_j(Z_i), \psi\}} \begin{pmatrix} \tilde{h}_{j\xi} \{G_j(Z_i), \tilde{\xi}, \psi\} \\ \tilde{h}_{j\theta} \{G_j(Z_i), \tilde{\xi}, \psi\} \end{pmatrix}, \\ \tilde{h}_{j\theta} \{G_j(Z_i), \tilde{\xi}, \psi\} &= -K_{j-1} \{G_j(Z_i), \psi\} \lambda_{j\psi} \{G_j(Z_i), \psi\} / \lambda_j \{G_j(Z_i), \psi\}, \\ \tilde{h}_{j\xi} \{G_j(Z_i), \tilde{\xi}, \psi\} &= h_{j\xi} \{G_j(Z_i), \xi, \psi\}. \end{aligned}$$

The matrix A_i is given by

$$A_i = \begin{pmatrix} A_{1i} \\ A_{2i} \end{pmatrix}, \quad A_{1i} = \begin{pmatrix} D_{i,1} & 0 & \cdots & \cdots & \cdots & 0 \\ 0 & \ddots & 0 & \cdots & \cdots & 0 \\ 0 & 0 & D_{i,r} & 0 & \cdots & 0 \\ 0 & \cdots & 0 & \ddots & 0 & 0 \\ 0 & \cdots & \cdots & 0 & D_{i,M} & 0 \end{pmatrix},$$

and

$$A_{2i} = \left(\partial / \partial \tau \{ \rho_{3,i}(\tau) \} \right),$$

where

$$D_{i,r} = -I(C_i \geq r) \lambda_r \{G_r(Z_i), \psi_r\} [1 - \lambda_r \{G_r(Z_i), \psi_r\}] \tilde{X}_{i,r}^T \tilde{X}_{i,r}, \quad r = 1, \dots, M,$$

$$\rho_{3,i}(\tau) = \left(\begin{array}{c} \sum_{r=1}^M I(C_i > r) \tilde{q}_r \{G_r(Z_i), \tilde{\xi}, \psi\} [\tilde{h}_{r+1} \{G_{r+1}(Z_i), \tilde{\xi}, \psi\} - \tilde{h}_r \{G_r(Z_i), \tilde{\xi}, \psi\}] \\ \frac{I(C_i = \infty) m(Z_i, \beta)}{\pi(\infty, Z_i, \psi)} + \sum_{r=1}^M \frac{dM_c \{r, G_r(Z_i), \psi\}}{K_r \{G_r(Z_i), \psi\}} h_r \{G_r(Z_i), \xi\} \end{array} \right).$$

Analogous to the strategy for $\hat{\beta}_{br*}$, in our implementation, we used numerical derivatives as an approximation to the analytical derivatives of $\rho_{3,i}(\tau)$ with respect to τ .

2.8.2 Derivation of conditional expectations implied by assumed mixed models in Section 2.5

We derive the required conditional expectations $E(Y|Y_1, \dots, Y_j, \tilde{X})$ for $j = 1, \dots, 4$ implied by model (17) in Section 2.5. The random vector $\Psi = (\alpha_0, \alpha_1, e_1, e_2, e_3, e_4)^T$ has multivariate normal distribution with mean μ and variance Σ , where

$$\mu = (\mu_{\alpha 0}, \mu_{\alpha 1}, 0_{1 \times 4})^T, \quad \Sigma = \begin{pmatrix} \Sigma_{\alpha} & 0_{2 \times 4} \\ 0_{4 \times 2} & \sigma_e^2 I_4 \end{pmatrix},$$

$0_{a \times b}$ is a zero matrix with dimension $(a \times b)$, and I_a is an $(a \times a)$ identity matrix. Therefore, the distribution of $(\alpha_0, \alpha_1, Y_1, Y_2, Y_3, Y_4)^T$, conditional on \tilde{X} , follows multivariate normal distribution with mean $\tilde{\mu} = A\mu + c$ and variance $\tilde{\Sigma} = A\Sigma A^T$, where

$$A = I_6 + \begin{pmatrix} 0_{2 \times 2} & 0_{2 \times 4} \\ A_{21} & 0_{4 \times 4} \end{pmatrix}, \quad A_{21} = \begin{pmatrix} 1 & 1 & 1 & 1 \\ t_1 & t_2 & t_3 & t_4 \end{pmatrix}^T, \quad \text{and} \quad c = \gamma^T \tilde{X} \begin{pmatrix} 0_{2 \times 1} \\ 1_{4 \times 1} \end{pmatrix}.$$

Hence, the conditional mean is given by

$$\begin{aligned} E(Y|Y_1, Y_2, Y_3, Y_4, \tilde{X}) &= E(Y_5|Y_1, Y_2, Y_3, Y_4, \tilde{X}) \\ &= \gamma^T X + E(\alpha_0|Y_1, Y_2, Y_3, Y_4, \tilde{X}) + t_5 E(\alpha_1|Y_1, Y_2, Y_3, Y_4, \tilde{X}). \end{aligned}$$

To calculate the conditional mean $E(\alpha_k|Y_1, Y_2, Y_3, Y_4, \tilde{X})$, $k = 0, 1$, we use the following property of multivariate normal distribution. Suppose $(X_1^T, X_2^T)^T$ follows a $N(v, \Omega)$ distribution. If v and Ω are partitioned correspondingly as follows:

$$v = \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} \quad \text{and} \quad \Omega = \begin{pmatrix} \Omega_{11} & \Omega_{12} \\ \Omega_{21} & \Omega_{22} \end{pmatrix},$$

then $(X_1|X_2 = a) \sim N(\bar{v}, \bar{\Omega})$, where $\bar{v} = v_1 + \Omega_{12}\Omega_{22}^{-1}(a - v_2)$. Straightforward application of the above property yields

$$E\left\{(\alpha_0, \alpha_1)^T|Y_1, \dots, Y_4, \tilde{X}\right\} = \tilde{\mu}_{1:2} + \tilde{\Sigma}_{1:2,3:6}\tilde{\Sigma}_{3:6,3:6}^{-1}\left\{(Y_1, Y_2, Y_3, Y_4)^T - \tilde{\mu}_{3:6}\right\},$$

where $\tilde{\mu}_{a:b}$ is a column vector consisting of a th to b th entries of $\tilde{\mu}$, and $\tilde{\Sigma}_{a:b,m:n}$ is the submatrix of $\tilde{\Sigma}$ with rows a to b and columns m to n . Therefore the conditional expectation is

$$E(Y|Y_1, Y_2, Y_3, Y_4, \tilde{X}) = \gamma^T \tilde{X} + (1, t_5) \left[\tilde{\mu}_{1:2} + \tilde{\Sigma}_{1:2,3:6}\tilde{\Sigma}_{3:6,3:6}^{-1} \left\{ (Y_1, Y_2, Y_3, Y_4)^T - \tilde{\mu}_{3:6} \right\} \right].$$

Similarly,

$$\begin{aligned} E(Y|Y_1, Y_2, Y_3, \tilde{X}) &= \gamma^T \tilde{X} + (1, t_5) \left[\tilde{\mu}_{1:2} + \tilde{\Sigma}_{1:2,3:5}\tilde{\Sigma}_{3:5,3:5}^{-1} \left\{ (Y_1, Y_2, Y_3)^T - \tilde{\mu}_{3:5} \right\} \right], \\ E(Y|Y_1, Y_2, \tilde{X}) &= \gamma^T \tilde{X} + (1, t_5) \left[\tilde{\mu}_{1:2} + \tilde{\Sigma}_{1:2,3:4}\tilde{\Sigma}_{3:4,3:4}^{-1} \left\{ (Y_1, Y_2)^T - \tilde{\mu}_{3:4} \right\} \right], \\ E(Y|Y_1, \tilde{X}) &= \gamma^T \tilde{X} + (1, t_5) \left\{ \tilde{\mu}_{1:2} + \tilde{\Sigma}_{1:2,3:3}\tilde{\Sigma}_{3:3,3:3}^{-1} (Y_1 - \tilde{\mu}_{3:3}) \right\}. \end{aligned}$$

Next, we provide the derivation of the conditional expectations

$$E(Y|Y_1, \dots, Y_j, \tilde{X}, \text{dis}_1, \text{dis}_2, \text{dis}_3, \text{dis}_4)$$

for $j = 1, \dots, 4$ implied by assumed linear mixed model used in the second, general coarsened data analysis in Section 2.5; i.e., we assumed that, for $r = 1, \dots, 5$, the data follow the linear mixed model

$$Y_{ir} = \alpha_{0i} + \alpha_{1i}t_{ir} + \gamma^T \tilde{X}_i + \phi_1 I(r \geq 3)\text{dis}_{i2} + \phi_2 I(r = 5)\text{dis}_{i4} + e_{ir},$$

where the random effects and within-subject deviations are normal as above, and now \tilde{X} consists of weight, karnof, and symp.

Following the same logic as above, the distribution of $(\alpha_0, \alpha_1, Y_1, Y_2, Y_3, Y_4)^T$, conditional on $(\tilde{X}, \text{dis}_1, \text{dis}_2, \text{dis}_3, \text{dis}_4)$, follows multivariate normal distribution with mean $\tilde{\mu}^* = A\mu + \tilde{c}$ and variance $\tilde{\Sigma} = A\Sigma A^T$, where $A, \mu, \Sigma, \tilde{\Sigma}$ are the same as above, and

$$\tilde{c} = \left(0_{1 \times 2}, \gamma^T \tilde{X}, \gamma^T \tilde{X}, \gamma^T \tilde{X} + \phi_1 \text{dis}_2, \gamma^T \tilde{X} + \phi_1 \text{dis}_2 \right)^T.$$

The conditional expectations are given as follows:

$$\begin{aligned} E(Y|Y_1, Y_2, Y_3, Y_4, \tilde{X}, \text{dis}_1, \text{dis}_2, \text{dis}_3, \text{dis}_4) \\ &= \gamma^T \tilde{X} + \phi_1 \text{dis}_2 + \phi_2 \text{dis}_4 + (1, t_5) \left[\tilde{\mu}_{1:2}^* + \tilde{\Sigma}_{1:2,3:6} \tilde{\Sigma}_{3:6,3:6}^{-1} \{ (Y_1, Y_2, Y_3, Y_4)^T - \tilde{\mu}_{3:6}^* \} \right], \\ E(Y|Y_1, Y_2, Y_3, \tilde{X}, \text{dis}_1, \text{dis}_2, \text{dis}_3, \text{dis}_4) \\ &= \gamma^T \tilde{X} + \phi_1 \text{dis}_2 + \phi_2 \text{dis}_4 + (1, t_5) \left[\tilde{\mu}_{1:2}^* + \tilde{\Sigma}_{1:2,3:5} \tilde{\Sigma}_{3:5,3:5}^{-1} \{ (Y_1, Y_2, Y_3)^T - \tilde{\mu}_{3:5}^* \} \right], \\ E(Y|Y_1, Y_2, \tilde{X}, \text{dis}_1, \text{dis}_2, \text{dis}_3, \text{dis}_4) \\ &= \gamma^T \tilde{X} + \phi_1 \text{dis}_2 + \phi_2 \text{dis}_4 + (1, t_5) \left[\tilde{\mu}_{1:2}^* + \tilde{\Sigma}_{1:2,3:4} \tilde{\Sigma}_{3:4,3:4}^{-1} \{ (Y_1, Y_2)^T - \tilde{\mu}_{3:4}^* \} \right], \\ E(Y|Y_1, \tilde{X}, \text{dis}_1, \text{dis}_2, \text{dis}_3, \text{dis}_4) \\ &= \gamma^T \tilde{X} + \phi_1 \text{dis}_2 + \phi_2 \text{dis}_4 + (1, t_5) \left\{ \tilde{\mu}_{1:2}^* + \tilde{\Sigma}_{1:2,3:3} \tilde{\Sigma}_{3:3,3:3}^{-1} (Y_1 - \tilde{\mu}_{3:3}^*) \right\}. \end{aligned}$$

2.8.3 Derivation of conditional expectations implied by assumed mixed model in Section 2.6

We derive the required conditional expectations $E(Y|\bar{L}_j)$ for $j = 1, 2$ implied by the model used in Section 2.6. The model implies that, in truth,

$$E(Y|\bar{L}_2) = E\{E(Y|\bar{L}_2, \alpha_0, \alpha_1)|\bar{L}_2\} = \gamma^T X + \mu_1(X, Y_1, Y_2) + t_3 \mu_2(X, Y_1, Y_2),$$

where $\mu_1(X, Y_1, Y_2) = E(\alpha_0|X, Y_1, Y_2)$, and $\mu_2(X, Y_1, Y_2) = E(\alpha_1|X, Y_1, Y_2)$.

Thus, we need to calculate the conditional distribution of α_0, α_1 given X, Y_1, Y_2 . The joint density of $(\alpha_0, \alpha_1, X, Y_1, Y_2)^T$ is given by

$$f(\alpha_0, \alpha_1, X, Y_1, Y_2) = f(Y_2|\alpha_0, \alpha_1, X, Y_1)f(Y_1|\alpha_0, \alpha_1, X)f(X)f(\alpha_0, \alpha_1)$$

Therefore,

$$\begin{aligned} f(\alpha_0, \alpha_1 | X, Y_1, Y_2) &= \frac{f(\alpha_0, \alpha_1, X, Y_1, Y_2)}{\int f(\alpha_0, \alpha_1, X, Y_1, Y_2) d\alpha_0 d\alpha_1} \\ &= \frac{f(Y_2 | \alpha_0, \alpha_1, X, Y_1) f(Y_1 | \alpha_0, \alpha_1, X) f(\alpha_0, \alpha_1)}{\int f(Y_2 | \alpha_0, \alpha_1, X, Y_1) f(Y_1 | \alpha_0, \alpha_1, X) f(\alpha_0, \alpha_1) d\alpha_0 d\alpha_1}. \end{aligned}$$

As a consequence,

$$f(\alpha_0, \alpha_1 | X, Y_1, Y_2) \propto f(Y_2 | \alpha_0, \alpha_1, X, Y_1) f(Y_1 | \alpha_0, \alpha_1, X) f(\alpha_0, \alpha_1).$$

After some algebra, it can be shown that, if we let $a = \sigma_{22}/(\sigma_{11}\sigma_{22} - \sigma_{12}^2)$, $b = -\sigma_{12}/(\sigma_{11}\sigma_{22} - \sigma_{12}^2)$, $c = \sigma_{11}/(\sigma_{11}\sigma_{22} - \sigma_{12}^2)$, $g_1(X, Y_1, Y_2) = a\mu_{\alpha_0} + b\mu_{\alpha_1} + (Y_2 + Y_1 - 2\gamma^T X)/\sigma_e^2$, and $g_2(X, Y_1, Y_2) = b\mu_{\alpha_0} + c\mu_{\alpha_1} + (Y_2 - \gamma^T X)/\sigma_e^2$, then

$$\begin{aligned} \mu_2(X, Y_1, Y_2) &= E(\alpha_1 | Z, Y_1, Y_2) = \frac{g_1(X, Y_1, Y_2) (b + 1/\sigma_e^2) - g_2(X, Y_1, Y_2) (a + 2/\sigma_e^2)}{(b + 1/\sigma_e^2)^2 - (c + 1/\sigma_e^2) (a + 2/\sigma_e^2)}, \\ \mu_1(X, Y_1, Y_2) &= E(\alpha_0 | Z, Y_1, Y_2) = \frac{g_2(X, Y_1, Y_2) - \mu_2(X, Y_1, Y_2) (c + 1/\sigma_e^2)}{b + 1/\sigma_e^2}. \end{aligned}$$

Similarly, we have

$$E(Y | \bar{L}_1) = E\{E(Y | \bar{L}_1, \alpha_0, \alpha_1) | \bar{L}_1\} = \gamma^T X + \mu_3(X, Y_1) + t_3 \mu_4(X, Y_1),$$

where $\mu_3(X, Y_1) = E(\alpha_0 | X, Y_1)$, and $\mu_4(X, Y_1) = E(\alpha_1 | X, Y_1)$. Letting $d = b\mu_{\alpha_0} + c\mu_{\alpha_1}$, $g_3(X, Y_1) = a\mu_{\alpha_0} + b\mu_{\alpha_1} + (Y_1 - \gamma^T X)/\sigma_e^2$, we have

$$\begin{aligned} \mu_3(X, Y_1) &= E(\alpha_0 | X, Y_1) = \frac{g_3(X, Y_1) \cdot c - d \cdot b}{(a + 1/\sigma_e^2) c - b^2}, \\ \mu_4(X, Y_1) &= E(\alpha_1 | X, Y_1) = \frac{d - \mu_3(X, Y_1) \cdot b}{c}. \end{aligned}$$

Bibliography

- [1] Bang, H. and Robins, J. M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics* **61**, 962–72.
- [2] Birmingham, J., Rotnitzky, A., and Fitzmaurice G.M. (2003). Pattern-mixture and selection models for analysing longitudinal data with monotone missing patterns. *Journal of the Royal Statistical Society, Series B* **65**, 275–297.
- [3] Davidian, M., Tsiatis, A. A., and Leon, S. (2005). Semiparametric estimation of treatment effect in a pretest-posttest study with missing data (with discussion and rejoinder). *Statistical Science* **20**, 261–301.
- [4] Gill, R. D., van der Laan, M. J., and Robins, J.M. (1997). Coarsening at random: characterizations, conjectures and counterexamples. *Proceedings of The First Seattle Symposium in Biostatistics: Survival Analysis*. New York: Springer, pp. 255–294.
- [5] Hammer, S. M., Katzenstein, D. A., Hughes, M. D., Gundaker, H., Schooley, R. T., Haubrich, R. H., Henry, W. K., Lederman, M. M., Phair, J. P., Niu, M., Hirsch, M. S., and Merigan, T. C., for the AIDS Clinical Trials Group Study 175 Study Team. (1996). A trial comparing nucleoside monotherapy with combination therapy in HIV infected adults with CD4 cell counts from 200 to 500 per cubic millimeter. *New England Journal of Medicine* **335**, 1081C1089.
- [6] Heitjan, D. F. (1993). Ignorability and coarse data: some biomedical examples. *Biometrics* **49**, 1099–1109.
- [7] Heitjan, D. F. and Rubin, D. B. (1991). Ignorability and coarse data. *The Annals of Statistics* **19**, 2244–2253.
- [8] Hogan, J. W., Roy, J., and Korkontzelou, C. (2004). Handling drop-out in longitudinal studies. *Statistics in Medicine* **23**, 1455–1497.
- [9] Kang, D. Y. and Schafer, J. L. (2007). Demystifying double robustness: a comparison of alternative strategies for estimating a population mean from incomplete data (with discussion and rejoinder). *Statistical Science* **22**, 523–380.
- [10] Little, R. J. A. (2009). Selection and pattern-mixture models. In *Longitudinal Data Analysis*, G. Fitzmaurice, M. Davidian, G. Verbeke, and G. Molenberghs (eds., 409–431. Boca Raton, FL: Chapman and Hall/CRC.
- [11] Lunceford, J. K. and Davidian, M. (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Statistics in Medicine* **23**, 2937–2960.

- [12] Molenberghs, G. and Fitzmaurice, G. (2009). Incomplete data: Introduction and overview. In *Longitudinal Data Analysis*, G. Fitzmaurice, M. Davidian, G. Verbeke, and G. Molenberghs (eds.), 395–408. Boca Raton, FL: Chapman and Hall/CRC.
- [13] Philipson, P. M., Ho, W. K., and Henderson, R. (2008). Comparative review of methods for handling drop-out in longitudinal studies. *Statistics in Medicine* **27**, 6276–6298.
- [14] Robins, J. M. (2000). Robust estimation in sequentially ignorable missing data and causal inference models. *Proceedings of the American Statistical Association Section on Bayesian Statistical Science 1999*, 6–10.
- [15] Robins, J. M., Hernán, M., and Brumback, B. (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology* **11**, 550–560.
- [16] Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association* **89**, 846–866.
- [17] Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association* **90**, 106–121.
- [18] Robins, J. M., Sued, M., Lei-Gomez, Q., and Rotnitzky, A. (2007). Performance of double-robust estimators when inverse probability weights are highly variable. *Statistical Science* **22**, 544–559.
- [19] Rosenbaum, P. R. (1987). Model-based direct adjustment. *Journal of the American Statistical Association* **82**, 387–394.
- [20] Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* **70**, 41–55.
- [21] Rosenbaum, P. R. and Rubin, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association* **79**, 516–524.
- [22] Rotnitzky, A. (2009). Inverse probability weighted methods. In *Longitudinal Data Analysis*, G. Fitzmaurice, M. Davidian, G. Verbeke, and G. Molenberghs (eds.), 453–476. Boca Raton, FL: Chapman and Hall/CRC.
- [23] Rotnitzky, A., Robins, J. M., and Scharfstein, D. O. (1998). Semiparametric regression for repeated outcomes with nonignorable nonresponse. *Journal of the American Statistical Association* **93**, 1321–1339.
- [24] Rubin, D. B. (1976). Inference and missing data. *Biometrika* **63**, 581–592.
- [25] Rubin, D. B. (1978). Bayesian inference for causal effects: the role of randomization. *The Annals of Statistics* **6**, 34–58.
- [26] Rubin, D. B. and Thomas N. (1996). Matching using estimated propensity scores: relating theory to practice. *Biometrics* **52**, 249–264.
- [27] SAS Institute, Inc. (2006). *SAS Online Documentation 9.1.3*. Cary, NC: SAS Institute.
- [28] Scharfstein, D. O., Rotnitzky, A., and Robins, J. M. (1999). Adjusting for nonignorable drop-out using semiparametric nonresponse models. (with discussion and rejoinder). *Journal of the American Statistical Association* **94**, 1096–1146.

- [29] Seaman, S. and Copas, A. (2009). Doubly robust generalized estimating equations for longitudinal data. *Statistics in Medicine* **28**, 937–955.
- [30] Stefanski, L. A. and Boos, D. D. (2002). The calculus of M-estimation. *The American Statistician* **56**, 29–38.
- [31] Tan, Z. (2006). A distributional approach for causal inference using propensity scores. *Journal of the American Statistical Association* **101**, 1619–1637.
- [32] Tan, Z. (2007). Understanding OR, PS and DR. *Statistical Science* **22**, 560–568.
- [33] Tan, Z. (2008). Comment: Improved Local Efficiency and Double Robustness. *The International Journal of Biostatistics* **4**, Issue 1, Article 10. doi: 10.2202/1557-4679.1109.
- [34] Tsiatis, A. A. (2006). *Semiparametric Theory and Missing Data*. New York: Springer.
- [35] Tsiatis, A. A. and Davidian, M. (2007). Comment on “Demystifying Double Robustness: A Comparison of Alternative Strategies for Estimating a Population Mean from Incomplete Data.” *Statistical Science* **22**, 569–573.

Appendices

SAS CODE FOR CHAPTER 1

Modeling missingness mechanism by enhanced propensity score model

```

%\begin{verbatim}
%let S=1000;
%let n=1000;

data enhanced (keep = z1 z2 z3 z4 y p R x1 x2 x3 x4 iter);
seed1 = 12 ;
seed2 = 123 ;
seed3 = 1234 ;
seed4 = 12345 ;
seed5=123456;
seed6=1234567;
do iter = 1 to &S;
do i=1 to &n;
z1=rannor(seed1);
z2=rannor(seed2);
z3=rannor(seed3);
z4=rannor(seed4);
y=210+27.4*z1+13.7*z2+13.7*z3+13.7*z4+rannor(seed5);
p=exp(-z1+0.5*z2-0.25*z3-0.1*z4)/(1+exp(-z1+0.5*z2-0.25*z3-0.1*z4));
R=ranbin(seed6,1,p);
y=y*R;
x1=exp(z1/2);
x2=z2/(1+exp(z1))+10;
x3=(z1*z3/25+0.6)**3;
x4=(z2+z4+20)**2;
output ;
end ;
end;
run;

PROC SORT ;
by iter;

```

```
proc logistic data=enhanced descending outest=parms noprint;
model R=z1 z2 z3 z4;
by iter;
output out=prob predicted=phat;
run;
```

```
/* parameters from the logistic regression */
data parms;
set parms;
keep Intercept z1 z2 z3 z4 iter;
run;
```

```
/* predicted values from logistic model fitting*/
data prob;
set prob;
keep iter phat;
run;
```

```
proc reg data=enhanced outest=lmparms noprint;
model y=z1 z2 z3 z4;
where R=1;
by iter;
run;
```

```
data lmparms;
set lmparms;
keep Intercept z1 z2 z3 z4 iter;
run;
```

```
proc iml;
a=j(&n,1,1.);
```

```
/* define P(R=1 | x) : enhanced ps model*/
start logistic(theta) global(z,a,z_new);
    templ=z_new * theta`;
```

```

        theta2=theta[2:6];
        temp2=z * theta2;
        f=1-exp(temp1)/(1+exp(temp2));
        f=f`;
        return(f);
finish logistic;

/* define the constraints */
start constrain(theta) global(z,a,z_new);
    temp1=z_new * theta`;
    theta2=theta[2:6];
    temp2=z * theta2;
    f=j(2*&n,1,.);
    f[1:&n,]=1-exp(temp1)/(1+exp(temp2))-0.02;
    f[(&n+1):(&n*2),]=exp(temp1)/(1+exp(temp2))-0.02;
    f=f`;
    return(f);
finish constrain;

/* define the likelihood to be maximized */
start likelh(theta) global(z,R,a,z_new);
    temp=logistic(theta);
    f=(log(temp))*R+(log(1-temp))*(1-R);
    return(f);
finish likelh;

/*derivatives of the log-likelihood (i.e., the score vector) to be used
in the optimazation procedure*/
start glike(theta) global(z,R,a,z_new);
    theta2=theta[2:6];
    temp=logistic(theta);
    g=j(1,6,.);
    g[1]=-1/temp *R+&n;
    temp2=z * theta2;
    denom=(1+exp(temp2`))#temp;

```

```

        g[2]=-(z[,1]`/denom)*(R-temp`);
        g[3]=-(z[,2]`/denom)*(R-temp`);
        g[4]=-(z[,3]`/denom)*(R-temp`);
        g[5]=-(z[,4]`/denom)*(R-temp`);
        g[6]=-(z[,5]`/denom)*(R-temp`);

    return(g);
finish glike;

/* Initialize a dataset to store: simulation number, parameter estimates
   and fitted values. */
sdata=j(&S, &n+7, 0.) ;
index=0;
muhat=0;
muhat_tan=0;
muhat_proj=0;
sddr=0;
sdtan=0;
sdproj=0;
flag_dr=0;
flag_tan=0;
flag_proj=0;

do i=1 to &S;

    /* read in the ith simulation data */
    use enhanced;
    read all Var{iter z1 z2 z3 z4 y p R x1 x2 x3 x4} where (iter=i)
    into fullmat;

    z=j(&n,5,0);
    z[,1]=a;
    z[,2:5]=fullmat[,2:5];
    x=j(&n,5,0);
    x[,1]=a;
    x[,2:5]=fullmat[,9:12];

```

```

R=fullmat[,8];
z_new=a||z;
y=fullmat[,6];

/* initial values for nlpqn */
x0=j(6,1,0);
use parms;
read all var{iter Intercept z1 z2 z3 z4} where (iter=i) into est
_logit;
est_mle=est_logit[2:6];

x0[2:6]=-est_mle;
epsilon={0,0,0,0,0,0};
x0=x0+epsilon;

optn=j(1,10,.); optn[1]=1; optn[2]=0; optn[10]=2*&n;

call nlpqn(rc,xres,"likelh",x0,optn,,,,,"glike") nlc="constrain
";
xopt1=xres`;
xpt=x0||xopt1;
*print xpt;

index=index || i;
sdata[i,1]=i;
sdata[i,2:7]=xres;

temp1_new=z_new * xres`;
theta2_new=xres[2:6];
temp2_new=z * theta2_new;
temp_new=1-exp(temp1_new)/(1+exp(temp2_new));
sdata[i,8:(&n+7)]=temp_new`;
pihat=sdata[i,8:(&n+7)];
/* read in the data from the partial lm fit and find predicted
values and residuals */

```

```

        use lmparms;
        read all var{iter Intercept z1 z2 z3 z4} where (iter=i)
        into est_lm;
est_lm=est_lm[2:6];

        /* fitted values and residuals */
fitted=z*est_lm;
        resid=y-fitted;
        mu_temp=sum(fitted)/&n + sum(R#resid/pihat`)/&n;
        muhat=muhat //mu_temp;

/*compute the standard error of DR estimator using emipirical
sandwich covariance matrix.
        form joint estimating equation with respect to gamma
        _0, gamma_1, beta and mu.
        var=inv(A_n)*B_n*inv(A_n)^T/n . */

/*First we compute the matrix B_n by defining the m-estimating
equation and then taking the product */
        B_n=j(12,12,0);
        do s=1 to &n;
                m_temp=j(12,1,0);
                m_temp[1]=(R[s]-pihat[s])/pihat[s];
                m_temp[2:6]=m_temp[1]#z[s,]`/(1+exp(temp2_new
[s]));
                m_temp[7:11]=R[s] # resid[s] # z[s,]`;
                m_temp[12]=R[s] # resid[s]/pihat[s]+fitted[s] - mu_temp;

                B_n=B_n+m_temp*m_temp`;
        end;
        B_n=B_n/&n;

/*Then we compute the matrix A_n by taking the derivatives of
the m-estimating equations. */
        s=1;
                A_n=j(12,12,0);

```

```

do s=1 to &n;
    a_temp=j(12,12,0);
    a_temp[1,1]=R[s]#(pihat[s]-1)/pihat[s]##2;
    a_temp[1,2:6]=a_temp[1]#z[s,]/(1+exp(temp2_new
    [s]));
    a_temp[2:6,1]=a_temp[1,2:6]`;
    a_temp[2:6,2:6]=(a_temp[1] + (R[s]-pihat[s])#exp(temp2_new
    [s])/pihat[s])/(1+exp(temp2_new[s]))##2 # z[s,]`*z[s,];
    a_temp[7:11,7:11]=R[s]#z[s,]`*z[s,];
    a_temp[12,1]=a_temp[1,1]#resid[s];
    a_temp[12,2:6]=a_temp[1,2:6]#resid[s];
    a_temp[12,7:11]=(R[s]-pihat[s])/pihat[s]#z[s
    ,];
    a_temp[12,12]=1;

    A_n=A_n+a_temp;
end;
A_n=A_n/&n;
A_inv=inv(A_n);
V_n=A_inv*B_n*A_inv`/&n;
sddr_temp=sqrt(V_n[12,12]);
sddr=sddr//sddr_temp;

ci_low=mu_temp-1.96*sddr_temp;
ci_up=mu_temp+1.96*sddr_temp;

if ci_low<210 & 210<ci_up then
    flag_dr=flag_dr+1;

/*Compute weights and do Tan's estimator*/
rindex=0;
do k=1 to &n;
    if R[k]>0 then
        rindex=rindex//k;
end;
rindex=rindex[2:nrow(rindex)];

```



```

        weights=(R-pihat`)/pihat`##2;
weights=weights[rindex];
/*specify matrix, and compute wls estimats */
weights=diag(weights);
        wz=z[rindex,];
        wy=y[rindex];
        est_tan=inv(wz`*weights*wz)*wz`*weights*wy;

fitted_tan=z*est_tan;
        resid_tan=y-fitted_tan;
        mu_temptan=sum(fitted_tan)/&n + sum(R#resid_tan/pihat
        `)/&n;
        muhat_tan=muhat_tan //mu_temptan;

/*First we compute the matrix B_n by defining the m-estimating
equation and then taking the product */
        B_tan=j(12,12,0);
s=1;
        weight_new=(R-pihat`)/pihat`##2;
do s=1 to &n;
        m_temp=j(12,1,0);
        m_temp[1]=(R[s]-pihat[s])/pihat[s];
        m_temp[2:6]=m_temp[1]*z[s,]`/(1+exp(temp2_new
        [s]));
        m_temp[7:11]=weight_new[s]* R[s] # resid_tan[s] # z[s
        ,]`;
        m_temp[12]=R[s] # resid_tan[s]/pihat[s]+fitted_tan[s] -
        mu_temptan;

        B_tan=B_tan+m_temp*m_temp`;
end;
        B_tan=B_tan/&n;

/*Then we compute the matrix A_n by taking the derivatives of
the m-estimating equations. */

```

```

s=1;
    A_tan=j(12,12,0);
    do s=1 to &n;
        a_temp=j(12,12,0);
        a_temp[1,1]=R[s]#(pihat[s]-1)/pihat[s]##2;
            a_temp[1,2:6]=a_temp[1]#z[s,]/(1+exp(temp2_new
                [s]));
            a_temp[2:6,1]=a_temp[1,2:6]`;
        a_temp[2:6,2:6]=(a_temp[1] + (R[s]-pihat[s])#exp(temp2_new
            [s])/pihat[s])/(1+exp(temp2_new[s]))##2 # z[s,]`*z[s,];

        a_temp[7:11,1]=a_temp[1,1]#(2-pihat[s])/pihat[s]#resid_tan
            [s]#z[s,]`;
            a_temp[7:11,2:6]=a_temp[1,1]#(2-pihat[s])/(pihat[s
                ]#(1+exp(temp2_new[s])))#resid_tan[s]#z[s,]`*z[s,];
        a_temp[7:11,7:11]=weight_new[s]#R[s]#z[s,]`*z[s,];

        a_temp[12,1]=a_temp[1,1]#resid_tan[s];
            a_temp[12,2:6]=a_temp[1,2:6]#resid_tan[s];
            a_temp[12,7:11]=(R[s]-pihat[s])/pihat[s]#z[s
                ,];
            a_temp[12,12]=1;

            A_tan=A_tan+a_temp;
    end;
    A_tan=A_tan/&n;
    A_taninv=inv(A_tan);
    V_tan=A_taninv*B_tan*A_taninv`/&n;
        sdtan_temp=sqrt(V_tan[12,12]);
    sdtan=sdtan//sdtan_temp;
    citan_low=mu_temptan-1.96*sdtan_temp;
        citan_up=mu_temptan+1.96*sdtan_temp;

    if citan_low<210 & 210<citan_up then
        flag_tan=flag_tan+1;

```

```

/* compute the new projection estimator */
/* define the big matrix [X, pi_gamma/(1-pi)] */
pgamma=j(&n,6,0);
do s=1 to &n;
    pgamma[s,1]=-1;
    pgamma[s,2:6]=-z[s,]/(1+exp(temp2_new[s]));
end;
newmat=z || pgamma;
wnewmat=newmat[rindex,];
    betac=ginv(wnewmat`*weights*wnewmat)*wnewmat`*weights*
    wy;
    beta=betac[1:5];
    estc=betac[6:11];
    fitted_proj=z*beta;
resid_proj=y-fitted_proj;
mu_tempproj=sum(fitted_proj)/&n + sum(R#resid_proj/pihat`)/&n -
    gluke(xres) * estc/&n ;
    muhat_proj=muhat_proj //mu_tempproj;

/* compute the standard error and coverage probability of the new
projection estimator */
/* define the matrix B_proj by taking the product of the M-vector
and sum them. */
B_proj=j(18,18,0);
    s=1;
do s=1 to &n;
    m_temp=j(18,1,0);
    m_temp[1]=(R[s]-pihat[s])/pihat[s];
    m_temp[2:6]=m_temp[1]#z[s,]/(1+exp(temp2_new[s]));
    m_temp[7:17]=weight_new[s] # R[s]# (resid_proj[s] - pgamma[s,]*
    estc) # newmat[s,];
* m_temp[18]= R[s]*resid_proj[s]/pihat[s] + fitted_proj[s] -
pgamma[s,]*estc # (R[s]-pihat[s])/(pihat[s]-pihat[s]##2)-mu_
tempproj;
    m_temp[18]= R[s]*resid_proj[s]/pihat[s] + fitted_proj[s] -

```

```

pgamma[s,]*estc # (R[s]-pihat[s])/pihat[s]-mu_tempproj;

B_proj=B_proj+m_temp*m_temp`;

end;

B_proj=B_proj/&n;

s=1;
  A_proj=j(18,18,0);
  do s=1 to &n;
    a_temp=j(18,18,0);
    a_temp[1,1]=R[s]#(pihat[s]-1)/pihat[s]##2;
    a_temp[1,2:6]=a_temp[1]#z[s,]/(1+exp(temp2_new[s]));
    a_temp[2:6,1]=a_temp[1,2:6]`;
    a_temp[2:6,2:6]=(a_temp[1] + (R[s]-pihat[s])#exp(temp2_new[s])/pihat[s])/(1+exp(temp2_new[s]))##2 # z[s,]*z[s,];
    a_temp[7:17,1]=R[s]#(pihat[s]-2#R[s])#(1-pihat[s])#(resid_proj[s] - pgamma[s,]*estc)#newmat[s,]`/pihat[s]##3;
    tempmat=j(6,5,0);
    tempmat[2:6,]=exp(temp2_new[s])/(1+exp(temp2_new[s]))##2 # z[s,]*z[s,];
    a_temp[7:11,2:6]=a_temp[7:11,1]*z[s,]/(1+exp(temp2_new[s]))+ R[s]#(R[s]-pihat[s])/pihat[s]##2 * z[s,]* estc` * tempmat;
    a_temp[7:11,7:11]=weight_new[s]#R[s]#z[s,]*z[s,];
    a_temp[7:11,12:17]=weight_new[s]#R[s]#z[s,]*pgamma[s,];

    t1=R[s]#(pihat[s]-2#R[s])#(1-pihat[s])#pgamma[s,]*z[s,]/((1+exp(temp2_new[s]))#pihat[s]##3)
    ;
    t2=R[s]#(R[s]-pihat[s])/pihat[s]##2# tempmat;
    t3=R[s]#(R[s]-pihat[s])/pihat[s]##2 # pgamma[s,]* estc`*tempmat;
    a_temp[12:17,2:6]= (t1-t2)#(resid_proj[s] - pgamma[s,]* estc) + t3;
    a_temp[12:17, 7:11]= R[s]#(R[s]-pihat[s])/pihat[s]##2 #

```

```

pgamma[s,]`*z[s,];
a_temp[12:17, 12:17]= R[s]#(R[s]-pihat[s])/
pihat[s]##2 # pgamma[s,]`*pgamma[s,];

a_temp[18,1]=R[s]#(pihat[s]-1)#resid_proj[s]/pihat[s]##2 -
pgamma[s,]*estc #R[s]#(pihat[s]-1)/pihat[s]##2;
a_temp[18,2:6]=a_temp[18,1]#z[s,]/(1+exp(temp2_new[s]))+ (
R[s]-pihat[s])/pihat[s]* estc`*tempmat;
a_temp[18,7:11]=(R[s]-pihat[s])/pihat[s]*z[s,];
a_temp[18,12:17]=(R[s]-pihat[s])/pihat[s]*pgamma[s,];
a_temp[18,18]=1;

A_proj=A_proj+a_temp;
end;
A_proj=A_proj/&n;
A_projinv=ginv(A_proj);
V_proj=A_projinv*B_proj*A_projinv`/&n;
sdproj_temp=sqrt(V_proj[18,18]);
sdproj=sdproj//sdproj_temp;
ciproj_low=mu_tempproj-1.96*sdproj_temp;
ciproj_up=mu_tempproj+1.96*sdproj_temp;

if ciproj_low<210 & 210<ciproj_up then
flag_proj=flag_proj+1;

terminate:
*print "invalid argument for log function";
aucillay=0;

end;

index=index[2:ncol(index)];
pp=sdata[index,];
/* estimated mean from DR estimator*/
muhat=muhat[2:nrow(muhat)];
muhat_tan=muhat_tan[2:nrow(muhat_tan)];
muhat_proj=muhat_proj[2:nrow(muhat_proj)];

```

```

*muhat=muhat[1:1000];
*muhat_tan=muhat_tan[1:1000];
sddr=sddr[2:nrow(sddr)];
sdtan=sdtan[2:nrow(sdtan)];
sdproj=sdproj[2:nrow(sdproj)];

mudata=muhat || muhat_tan || muhat_proj;
cname={"DR" "Tan" "proj"};
create outmu from mudata [colname=cname];
append from mudata;

sddata=sddr||sdtan || sdproj;
create outsd from sddata [colname=cname];
append from sddata;

cover_dr=flag_dr/1000;
cover_tan=flag_tan/1000;
cover_proj=flag_proj/1000;
print cover_dr;
print cover_tan;
print cover_proj;

quit;

proc univariate data=outmu;
run;

data outmu2;
set outmu;
absmuhat=abs(DR-210);
absmutan=abs(Tan-210);
absmuproj=abs(proj-210);
drop DR Tan proj;
run;

proc means data=outmu2 median;

```

```
run;  
  
proc means data=outsd mean;  
run;
```

SAS CODE FOR CHAPTER 2

```

/*****/
IPW and BR estimator

/*****/

ods listing close;
proc logistic data=actg2 descending;
model C1=cd40 age wtkg hemo homo  drugs  karnof  z30  preanti race
gender str2 symptom/selection=forward slentry=0.15;
output out=prob1 (drop=_level_) predicted=phat1;
ods output ParameterEstimates=parestc1;
run;
ods listing;

proc sql noprint;
select variable into :parc1 separated by ' '
from parestc1 where variable ne 'Intercept';
quit;
%put &parc1;

ods listing close;
proc logistic data=prob1 descending;
model C2=cd40 cd420 age wtkg hemo homo  drugs  karnof  z30  preanti race
gender str2 symptom /selection=forward slentry=0.15;
output out=prob2 (drop=_level_) predicted=phat2;
ods output ParameterEstimates=parestc2;
run;
ods listing;
proc sql noprint;
select variable into :parc2 separated by ' '
from parestc2 where variable ne 'Intercept';

```



```

quit;
%put &parc2;

ods listing close;
proc logistic data=prob2 descending;
model C3=cd40 cd420 cd440 age wtkg hemo homo  drugs  karnof  z30
preanti race gender str2 symptom/selection=forward slentry=0.15;
output out=prob3 (drop=__level__) predicted=phat3;
ods output ParameterEstimates=parestc3;
run;
ods listing;

proc sql noprint;
select variable into :parc3 separated by ' '
from parestc3 where variable ne 'Intercept';
quit;
%put &parc3;

ods listing close;
proc logistic data=prob3 descending ;
model C4=cd40 cd420 cd440 cd460 age wtkg hemo homo  drugs  karnof  z30
preanti race gender str2 symptom/selection=forward slentry=0.15;
output out=prob4 (drop=__level__) predicted=phat4;
ods output ParameterEstimates=parestc4;
run;
ods listing;

proc sql noprint;
select variable into :parc4 separated by ' '
from parestc4 where variable ne 'Intercept';
quit;
%put &parc4;

data IPW;
set prob4;

```

```

pi_5=(1-phat1)*(1-phat2)*(1-phat3)*(1-phat4);
ipi_5=1/pi_5;
pi_4=(1-phat1)*(1-phat2)*(1-phat3);
ipi_4=1/pi_4;
pi_3=(1-phat1)*(1-phat2);
ipi_3=1/pi_3;
pi_2=(1-phat1);
ipi_2=1/pi_2;
wcd496=cd496/pi_5;
run;

```

```

proc means data=IPW sum noprint;
var wcd496;
output out=ipwsum sum=wsum_cd496;
run;

```

```

data ipwresults;
set ipwsum;
cd496=500*wsum_cd496/_freq_;
run;

```

```

proc iml;

```

```

use ipwresults;
read all var{cd496} into mu_ipw;
mu_ipw=mu_ipw/500;
print mu_ipw;

```

```

use ipw;
read all var{cd40 cd420 cd440 cd460 cd496
C1 C2 C3 C4

```

```
/*1:5*/
```

```

phat1 phat2 phat3 phat4 pi_2 pi_3 pi_4 pi_5
/
dis1 dis2 dis4

```

```
/*6:9*/
```

```
/*10:17*
```

```

}                                /*25:31*/
into fulldata;

test=fulldata[1,];
print test;

y1=fulldata[,1];
y2=fulldata[,2];
y3=fulldata[,3];
y4=fulldata[,4];
y5=fulldata[,5];

/*missingness indicator*/
c1=fulldata[,6];
c2=fulldata[,7];
c3=fulldata[,8];
c4=fulldata[,9];
/*discrete and cumulative hazard function*/
lambda1=fulldata[,10];
lambda2=fulldata[,11];
lambda3=fulldata[,12];
lambda4=fulldata[,13];
k1=fulldata[,14];
k2=fulldata[,15];
k3=fulldata[,16];
k4=fulldata[,17];
dis1=fulldata[,18];
dis2=fulldata[,19];
dis4=fulldata[,20];

n=nrow(y1); /*number of observations*/
a=j(n,1,1);

/*X=fulldata[,21:24];
x2=fulldata[,25:ncol(fulldata)];

```

```

*/
use ipw;
read all var{&parc1} into x_lam1;
read all var{&parc2} into x_lam2;
read all var{&parc3} into x_lam3;
read all var{&parc4} into x_lam4;

x_lam1=a || x_lam1;
x_lam2=a || x_lam2;
x_lam3=a || x_lam3;
x_lam4=a || x_lam4;

p1=ncol(x_lam1);
p2=ncol(x_lam2);
p3=ncol(x_lam3);
p4=ncol(x_lam4);
tp=1+p1+p2+p3+p4;

print tp;

B_n=j(tp,tp,0);

do s=1 to n;

    m_temp=j(tp,1,0);
    m_temp[1:p1]=(c1[s]-lambda1[s])#x_lam1[s,]`;
    if c1[s]=0 then
        m_temp[(p1+1):(p1+p2)]=(c2[s]-lambda2[s])#x_
            lam2[s,]`;
    if c2[s]=0 then
        m_temp[(p1+p2+1):(p1+p2+p3)]=(c3[s]-lambda3[s])#x_lam3[s
            ,]`;
        if c3[s]=0 then
            m_temp[(p1+p2+p3+1):(tp-1)]=(c4[s]-lambda4[s
                ])#x_lam4[s,]`;

    m_temp[tp]=(c4[s]=0)*(y5[s]/k4[s]) - mu_ipw;

```

```
B_n=B_n+m_temp*m_temp`;
```

```
end;
```

```
B_n=B_n/n;
```

```
/*Then we compute the matrix A_n by taking the derivatives of  
the m-estimating equations. */
```

```
A_n=j(tp,tp,0);
```

```
do s=1 to n;
```

```
    a_temp=j(tp,tp,0);
```

```
    a_temp[1:p1,1:p1]=lambda1[s]#(1-lambda1[s])#x_lam1[s,]`*x_  
    lam1[s,];
```

```
    if c1[s]=0 then
```

```
    a_temp[(p1+1):(p1+p2),(p1+1):(p1+p2)]=lambda2[s]#(1-  
    lambda2[s])#x_lam2[s,]`*x_lam2[s,];
```

```
    if c2[s]=0 then
```

```
    a_temp[(p1+p2+1):(p1+p2+p3),(p1+p2+1):(p1+p2+p3)]=lambda3[  
    s]#(1-lambda3[s])#x_lam3[s,]`*x_lam3[s,];
```

```
    if c3[s]=0 then
```

```
    a_temp[(p1+p2+p3+1):(tp-1),(p1+p2+p3+1):(tp-1)]=lambda4[s  
    ]#(1-lambda4[s])#x_lam4[s,]`*x_lam4[s,];
```

```
if c4[s]=0 then
```

```
    do;
```

```
        a_temp[tp,1:p1]=-y5[s]*lambda1[s]/k4[s]*x_lam1[s,];
```

```
        a_temp[tp,(p1+1):(p1+p2)]=-y5[s]*lambda2[s]/k4[s]*x_  
        lam2[s,];;
```

```
        a_temp[tp,(p1+p2+1):(p1+p2+p3)]=-y5[s]*lambda3[s]/k4[  
        s]*x_lam3[s,];;
```

```
        a_temp[tp,(p1+p2+p3+1):(tp-1)]=-y5[s]*lambda4[s]/k4[s  
        ]*x_lam4[s,];;
```

```
    end;
```

```
        a_temp[tp,tp]=1;
```

```

                                A_n=A_n+a_temp;
                                end;
                                A_n=A_n/n;
                                A_inv=inv(A_n);
                                V_n=A_inv*B_n*A_inv\'/n;
                                sderr=500*sqrt(V_n[tp,tp]);

                                print sderr;

quit;

ods listing close;
proc reg data=ipw;
model cd496=cd40 cd420 cd440 cd460 wtkg karnof str2 symptom;
ods output ParameterEstimates=parm5;
run;
quit;
ods listing;
*ods trace off;

proc sql noprint;
select variable into :parm5 separated by ' '
from parm5 where variable ne 'Intercept';
quit;
%put &parm5;

ods listing close;
proc reg data=ipw outest=parest5 ;
model cd496=ipi_5 &parm5;
*where c3=0;
output out=ipw2 predicted=h4;
run;
quit;
ods listing;

```

```
ods listing close;
proc reg data=ipw2;
model h4=cd40 cd420 cd440 wtkg karnof str2 symptom ;
ods output ParameterEstimates=parm4;
run;
quit;
ods listing;
```

```
proc sql noprint;
select variable into :parm4 separated by ' '
from parm4 where variable ne 'Intercept';
quit;
%put &parm4;
```

```
ods listing close;
proc reg data=ipw2 outest=parest4;
model h4=ipi_4 &parm4;
*where c2=0;
output out=ipw3 predicted=h3;
run;
quit;
ods listing;
```

```
ods listing close;
proc reg data=ipw3;
model h3=cd40 cd420 wtkg karnof str2 symptom;
ods output ParameterEstimates=parm3;
run;
quit;
ods listing;
```

```
proc sql noprint;
select variable into :parm3 separated by ' '
from parm3 where variable ne 'Intercept';
```

```

quit;
%put &parm3;

ods listing close;
proc reg data=ipw3 outest=parest3;
model h3=ipi_3 &parm3;
*where c1=0;
output out=ipw4 predicted=h2;
run;
quit;
ods listing;

ods listing close;
proc reg data=ipw4;
model h2= cd40 wtkg karnof str2 symptom;
ods output ParameterEstimates=parm2;
run;
quit;
ods listing;

proc sql noprint;
select variable into :parm2 separated by ' '
from parm2 where variable ne 'Intercept';
quit;
%put &parm2;

ods listing close;
proc reg data=ipw4 outest=parest2;
model h2=ipi_2 &parm2;
output out=ipw5 predicted=h1;
run;
quit;
ods listing;

proc means data=ipw5 noprint;

```



```
var h1;
output out=meany mean=mu_br;
run;
```

```
data meany;
set meany;
mu_br=mu_br*500;
run;
```

```
proc print data=meany;
run;
```

```
proc iml;
```

```
use parest5;
read all var{ipi_5} into coeff5;
use parest4;
read all var{ipi_4} into coeff4;
use parest3;
read all var{ipi_3} into coeff3;
use parest2;
read all var{ipi_2} into coeff2;
```

```
use meany;
read all var{mu_br} into mu_br;
mu_br=mu_br/500;
```

```
use ipw5;
read all var{cd40 cd420 cd440 cd460 cd496
C1 C2 C3 C4
```

```
/*1:5*/
```

```
/*6:9*/
```

```
phat1 phat2 phat3 phat4 pi_2 pi_3 pi_4 pi_5
/
```

```
/*10:17*
```

```
ipi_5 ipi_4 ipi_3 ipi_2
/
```

```
/*18:21*
```

```

dis1 dis2 dis3 dis4                                /*22:25*/
h4 h3 h2 h1                                         /*26:29*/
} into fulldata;

y1=fulldata[,1]; y2=fulldata[,2]; y3=fulldata[,3]; y4=fulldata[,4]; y5=
fulldata[,5];

n=nrow(y1); /*number of observations*/
a=j(n,1,1);
/*missingness indicator*/
c1=fulldata[,6]; c2=fulldata[,7]; c3=fulldata[,8]; c4=fulldata[,9];
/*discrete and cumulative hazard function*/
lambda1=fulldata[,10]; lambda2=fulldata[,11]; lambda3=fulldata[,12];
lambda4=fulldata[,13];
k1=fulldata[,14]; k2=fulldata[,15]; k3=fulldata[,16]; k4=fulldata[,17];
ipi_5=fulldata[,18]; ipi_4=fulldata[,19]; ipi_3=fulldata[,20]; ipi_2=
fulldata[,21];
h4=fulldata[,26]; h3=fulldata[,27]; h2=fulldata[,28]; h1=fulldata[,29];

use ipw5;
read all var{ &parc1 } into x1;
read all var{ &parc2 } into x2;
read all var{ &parc3 } into x3;
read all var{ &parc4 } into x4;
read all var{ &parm5 } into x5;
read all var{ &parm4 } into x6;
read all var{ &parm3 } into x7;
read all var{ &parm2 } into x8;

x1=a || x1; x2=a || x2;
x3=a || x3; x4=a || x4;

x5=a || ipi_5 || x5; x6=a || ipi_4 || x6;
x7=a || ipi_3 || x7; x8=a || ipi_2 || x8;

p1=ncol(x1); p2=ncol(x2); p3=ncol(x3); p4=ncol(x4);

```

```

p5=ncol(x5); p6=ncol(x6); p7=ncol(x7); p8=ncol(x8);
tp=1+p1+p2+p3+p4+p5+p6+p7+p8;

print tp;
B_n=j(tp,tp,0);

do s=1 to n;

    m_temp=j(tp,1,0);
    m_temp[1:p1]=(c1[s]-lambda1[s])#x1[s,]`;
    if c1[s]=0 then
        m_temp[(p1+1):(p1+p2)]=(c2[s]-lambda2[s])#x2[
            s,]`;
    if c2[s]=0 then
        m_temp[(p1+p2+1):(p1+p2+p3)]=(c3[s]-lambda3[s])#x3[s,]`;
        if c3[s]=0 then
            m_temp[(p1+p2+p3+1):(p1+p2+p3+p4)]=(c4[s]-
                lambda4[s])#x4[s,]`;

    if c4[s]=0 then
        m_temp[(p1+p2+p3+p4+1):(p1+p2+p3+p4+p5)]=(y5[s]-h4[s])#x5
            [s,]`;
    if c3[s]=0 then
        m_temp[(p1+p2+p3+p4+p5+1):(p1+p2+p3+p4+p5+p6)]=(h4[s]-h3[
            s])#x6[s,]`;
        if c2[s]=0 then
            m_temp[(p1+p2+p3+p4+p5+p6+1):(p1+p2+p3+p4+p5+p6+p7)]=(h3[
                s]-h2[s])#x7[s,]`;
    if c1[s]=0 then
        m_temp[(p1+p2+p3+p4+p5+p6+p7+1):(p1+p2+p3+p4+p5+p6+p7+p8)
            ]=(h2[s]-h1[s])#x8[s,]`;

    m_temp[tp]=h1[s]-mu_br;

    B_n=B_n+m_temp*m_temp`;
end;

B_n=B_n/n;

```

/*Then we compute the matrix A_n by taking the derivatives of the m-estimating equations. */

```

A_n=j(tp,tp,0);
do s=1 to n;
    a_temp=j(tp,tp,0);
    a_temp[(1:p1),(1:p1)]=lambda1[s]#(1-lambda1[s])#x1[s,]`*x1
[s,];
    if c1[s]=0 then
a_temp[(p1+1):(p1+p2),(p1+1):(p1+p2)]=lambda2[s]#(1-
lambda2[s])#x2[s,]`*x2[s,];
    if c2[s]=0 then
a_temp[(p1+p2+1):(p1+p2+p3),(p1+p2+1):(p1+p2+p3)]=lambda3[
s]#(1-lambda3[s])#x3[s,]`*x3[s,];
    if c3[s]=0 then
a_temp[(p1+p2+p3+1):(p1+p2+p3+p4),(p1+p2+p3+1):(p1+p2+p3+
p4)]=lambda4[s]#(1-lambda4[s])#x4[s,]`*x4[s,];

if c4[s]=0 then
    do;
        mat1=j(ncol(x5), ncol(x1),0);
        mat1[2,]=lambda1[s]*x1[s,]*ipi_5[s];

        mat2=j(ncol(x5), ncol(x2),0);
        mat2[2,]=lambda2[s]*x2[s,]*ipi_5[s];

        mat3=j(ncol(x5), ncol(x3),0);
        mat3[2,]=lambda3[s]*x3[s,]*ipi_5[s];

        mat4=j(ncol(x5), ncol(x4),0);
        mat4[2,]=lambda4[s]*x4[s,]*ipi_5[s];

        a_temp[(p1+p2+p3+p4+1):(p1+p2+p3+p4+

```

```

p5), (1:p1)]=lambda1[s]#coeff5#x5[s
,]`*x1[s,]*ipi_5[s]-(y5[s]-h4[s])*
mat1;
a_temp[(p1+p2+p3+p4+1):(p1+p2+p3+p4+p5), (p1+1):(p1+
p2)]=lambda2[s]*coeff5*x5[s,]`*x2[s,]*ipi_5[s]-(y5[s
]-h4[s])*mat2;
a_temp[(p1+p2+p3+p4+1):(p1+p2+p3+p4+
p5), (p1+p2+1):(p1+p2+p3)]=lambda3[s]
*coeff5*x5[s,]`*x3[s,]*ipi_5[s]-(y5[s
]-h4[s])*mat3;
a_temp[(p1+p2+p3+p4+1):(p1+p2+p3+p4+p5), (p1+p2+p3+1)
:(p1+p2+p3+p4)]=lambda4[s]*coeff5*x5[s,]`*x4[s,]*ipi_
5[s]-(y5[s]-h4[s])*mat4;
a_temp[(p1+p2+p3+p4+1):(p1+p2+p3+p4+p5), (p1+p2+p3+
p4+1):(p1+p2+p3+p4+p5)]=x5[s,]`*x5[s,];
end;

if c3[s]=0 then
do;
mat1=j(ncol(x6), ncol(x1),0);
mat1[2,]=lambda1[s]*x1[s,]*ipi_4[s];

mat2=j(ncol(x6), ncol(x2),0);
mat2[2,]=lambda2[s]*x2[s,]*ipi_4[s];

mat3=j(ncol(x6), ncol(x3),0);
mat3[2,]=lambda3[s]*x3[s,]*ipi_4[s];

/* mat4=j(ncol(x6), ncol(x4),0);
mat4[2,]=lambda4[s]*x4[s,]*ipi_4[s];
*/

a_temp[(p1+p2+p3+p4+p5+1):(p1+p2+p3+p4+p5+p6), (1:p1)
]=-x6[s,]`*x1[s,]*lambda1[s]*(coeff5*ipi_5[s]-coeff4*
ipi_4[s])-(h4[s]-h3[s])*mat1;
a_temp[(p1+p2+p3+p4+p5+1):(p1+p2+p3+p4+p5+p6), (p1+1)

```

```

: (p1+p2)] = -x6[s,] * x2[s,] * lambda2[s] * (coeff5*ipi_5[s]
]-coeff4*ipi_4[s]) - (h4[s]-h3[s]) * mat2;
a_temp[(p1+p2+p3+p4+p5+1):(p1+p2+p3+p4+p5+p6), (p1+p2
+1):(p1+p2+p3)] = -x6[s,] * x3[s,] * lambda3[s] * (coeff5*
ipi_5[s]-coeff4*ipi_4[s]) - (h4[s]-h3[s]) * mat3;
*   a_temp[(p1+p2+p3+p4+p5+1):(p1+p2+p3+p4+p5+p6), (p1+
p2+p3+1):(p1+p2+p3+p4)] = -x6[s,] * x4[s,] * lambda4[s] * (
coeff5*ipi_5[s]-coeff4*ipi_4[s]) - (h4[s]-h3[s]) * mat4;
    a_temp[(p1+p2+p3+p4+p5+1):(p1+p2+p3+p4+p5+p6), (p1+
p2+p3+1):(p1+p2+p3+p4)] = -x6[s,] * x4[s,] * lambda4[s] *
coeff5*ipi_5[s];
a_temp[(p1+p2+p3+p4+p5+1):(p1+p2+p3+p4+p5+p6), (p1+p2
+p3+p4+1):(p1+p2+p3+p4+p5)] = -x6[s,] * x5[s,];
a_temp[(p1+p2+p3+p4+p5+1):(p1+p2+p3+p4+p5+p6), (p1+p2
+p3+p4+p5+1):(p1+p2+p3+p4+p5+p6)] = x6[s,] * x6[s,];
    end;

    if c2[s]=0 then
    do;
mat1=j(ncol(x7), ncol(x1),0);
        mat1[2,]=lambda1[s]*x1[s,]*ipi_3[s];

mat2=j(ncol(x7), ncol(x2),0);
        mat2[2,]=lambda2[s]*x2[s,]*ipi_3[s];

/*      mat3=j(ncol(x7), ncol(x3),0);
        mat3[2,]=lambda3[s]*x3[s,]*ipi_3[s];
*/

a_temp[(p1+p2+p3+p4+p5+p6+1):(p1+p2+p3+p4+p5+p6+p7),
(1:p1)] = -x7[s,] * x1[s,] * lambda1[s] * (coeff4*ipi_4[s]-
coeff3*ipi_3[s]) - (h3[s]-h2[s]) * mat1;
a_temp[(p1+p2+p3+p4+p5+p6+1):(p1+p2+p3+p4+p5+p6+p7),
(p1+1):(p1+p2)] = -x7[s,] * x2[s,] * lambda2[s] * (coeff4*
ipi_4[s]-coeff3*ipi_3[s]) - (h3[s]-h2[s]) * mat2;
a_temp[(p1+p2+p3+p4+p5+p6+1):(p1+p2+p3+p4+p5+p6+p7),

```

```

(p1+p2+1):(p1+p2+p3)]=-x7[s,]`*x3[s,]*lambda3[s]*
coeff4*ipi_4[s];

a_temp[(p1+p2+p3+p4+p5+p6+1):(p1+p2+p3+p4+p5+p6+p7),
(p1+p2+p3+p4+p5+1):(p1+p2+p3+p4+p5+p6)]=-x7[s,]`*x6[s
,];
a_temp[(p1+p2+p3+p4+p5+p6+1):(p1+p2+p3+p4+p5+p6+p7),
(p1+p2+p3+p4+p5+p6+1):(p1+p2+p3+p4+p5+p6+p7)]=x7[s,]`
*x7[s,];
end;

if c1[s]=0 then
do;
mat1=j(ncol(x8), ncol(x1),0);
mat1[2,]=lambda1[s]*x1[s,]*ipi_2[s];

a_temp[(p1+p2+p3+p4+p5+p6+p7+1):(tp-1), (1:p1)]=-x8[s
,]`*x1[s,]*lambda1[s]*(coeff3*ipi_3[s]-coeff2*ipi_2[s
])-(h2[s]-h1[s])*mat1;
* a_temp[(p1+p2+p3+p4+p5+p6+p7+1):(tp-1), (p1+1):(p1+
p2)]=-x8[s,]`*x2[s,]*lambda2[s]*(coeff3*ipi_3[s]-
coeff2*ipi_2[s])-(h2[s]-h1[s])*mat2;
a_temp[(p1+p2+p3+p4+p5+p6+p7+1):(tp-1), (p1+1):(p1+p2
)]=-x8[s,]`*x2[s,]*lambda2[s]*coeff3*ipi_3[s];

a_temp[(p1+p2+p3+p4+p5+p6+p7+1):(tp-1), (p1+p2+p3+p4+
p5+p6+1):(p1+p2+p3+p4+p5+p6+p7)]=-x8[s,]`*x7[s,];
a_temp[(p1+p2+p3+p4+p5+p6+p7+1):(tp-1), (p1+p2+p3+p4+
p5+p6+p7+1):(tp-1)]=x8[s,]`*x8[s,];
end;

a_temp[tp, (1:p1)]=-x1[s,]*lambda1[s]*coeff2*ipi_2[s
];

a_temp[tp, (p1+p2+p3+p4+p5+p6+p7+1):(

```

```

                                tp-1) ]=-x8[s,];
                                a_temp[tp,tp]=1;

                                A_n=A_n+a_temp;
                                end;
                                * print A_n;

                                A_n=A_n/n;
                                A_inv=inv(A_n);
                                V_n=A_inv*B_n*A_inv\ /n;
                                sderr=500*sqrt(V_n[tp,tp]);

                                print sderr;

                                quit;

                                /*****/
                                Projection estimator

                                /*****/
                                ods listing close;
                                proc logistic data=actg2 descending;
                                model C1=cd40 age wtkg hemo homo  drugs  karnof  z30  preanti race
                                gender str2 symptom/selection=forward slentry=0.15;
                                output out=prob1 (drop=_level_) predicted=phat1;
                                ods output ParameterEstimates=parestc1;
                                run;
                                ods listing;

                                proc sql noprint;
                                select variable into :parc1 separated by ' '
                                from parestc1 where variable ne 'Intercept';

```



```

quit;
%put &parc1;

ods listing close;
proc logistic data=prob1 descending;
model C2=cd40 cd420 age wtkg hemo homo  drugs  karnof  z30  preanti race
      gender str2 symptom /selection=forward slentry=0.15;
output out=prob2 (drop=_level_) predicted=phat2;
ods output ParameterEstimates=parestc2;
run;
ods listing;

proc sql noprint;
select variable into :parc2 separated by ' '
from parestc2 where variable ne 'Intercept';
quit;
%put &parc2;

ods listing close;
proc logistic data=prob2 descending;
model C3=cd40 cd420 cd440 age wtkg hemo homo  drugs  karnof  z30
preanti race gender str2 symptom/selection=forward slentry=0.15;
output out=prob3 (drop=_level_) predicted=phat3;
ods output ParameterEstimates=parestc3;
run;
ods listing;

proc sql noprint;
select variable into :parc3 separated by ' '
from parestc3 where variable ne 'Intercept';
quit;
%put &parc3;

ods listing close;
proc logistic data=prob3 descending;
model C4=cd40 cd420 cd440 cd460 age wtkg hemo homo  drugs  karnof  z30

```

```

preanti race gender str2 symptom/selection=forward slentry=0.15;
output out=prob4 (drop=_level_) predicted=phat4;
ods output ParameterEstimates=parestc4;
run;
ods listing;

```

```

proc sql noprint;
select variable into :parc4 separated by ' '
from parestc4 where variable ne 'Intercept';
quit;
%put &parc4;

```

```

data IPW;
set prob4;
pi_5=(1-phat1)*(1-phat2)*(1-phat3)*(1-phat4);
ipi_5=1/pi_5;
pi_4=(1-phat1)*(1-phat2)*(1-phat3);
ipi_4=1/pi_4;
pi_3=(1-phat1)*(1-phat2);
ipi_3=1/pi_3;
pi_2=(1-phat1);
ipi_2=1/pi_2;

```

```

wcd496=cd496/pi_5;
run;

```

```

proc means data=IPW sum noprint;
var wcd496;
output out=ipwsum sum=wsum_cd496;
run;

```

```

data ipwresults;
set ipwsum;
cd496=500*wsum_cd496/_freq_;
run;

```

```
proc print data=ipwresults;
run;
```

```
proc iml symsize=82920000 WORKSIZE=8292000000;
```

```
use ipw;
read all var{cd40 cd420 cd440 cd460 cd496          /*1:5*/
C1 C2 C3 C4
                                     /*6:9*/
phat1 phat2 phat3 phat4 pi_2 pi_3 pi_4 pi_5      /*10:17*/
/
dis1 dis2 dis4
                                     /*18:20*/
wtkg karnof oprior symptom
                                     /*21:24*/
age drugs preanti str2 homo hemo race}          /*25:31*/
into fulldata;

y1=fulldata[,1];
y2=fulldata[,2];
y3=fulldata[,3];
y4=fulldata[,4];
y5=fulldata[,5];

/*missingness indicator*/
c1=fulldata[,6];
c2=fulldata[,7];
c3=fulldata[,8];
c4=fulldata[,9];
/*discrete and cumulative hazard function*/
lambda1=fulldata[,10];
lambda2=fulldata[,11];
lambda3=fulldata[,12];
lambda4=fulldata[,13];
```

```

k1=fulldata[,14];
k2=fulldata[,15];
k3=fulldata[,16];
k4=fulldata[,17];

n=nrow(y1); /*number of observations*/
a=j(n,1,1);

X=fulldata[,21:22] || fulldata[,28] || fulldata[,24];
p=ncol(X); /*number of columns for X, p=4*/

use ipw;
read all var{ &parc1 } into x_lam1;
read all var{ &parc2 } into x_lam2;
read all var{ &parc3 } into x_lam3;
read all var{ &parc4 } into x_lam4;

x_lam1=a || x_lam1; x_lam2=a || x_lam2;
x_lam3=a || x_lam3; x_lam4=a || x_lam4;

p1=ncol(x_lam1);
p2=ncol(x_lam2);
p3=ncol(x_lam3);
p4=ncol(x_lam4);

tp=6+p+p1+p2+p3+p4; /*mu_alpha0, mu_alpha1, sigma1^2, sigma2^2, sigma12
, sigma^2 */

time={0 1 2 3 4.8};

rindex2=0;

```

```

do i=1 to n;
    if c1[i]=0 then rindex2=rindex2//i;
end;
rindex2=rindex2[2:nrow(rindex2),];

rindex3=0;
do i=1 to n;
    if c2[i]=0 then rindex3=rindex3//i;
end;
rindex3=rindex3[2:nrow(rindex3),];

rindex4=0;
do i=1 to n;
    if c3[i]=0 then rindex4=rindex4//i;
end;
rindex4=rindex4[2:nrow(rindex4),];

y12= y1 || y2;
y123=y1 || y2 || y3;
y1234=y1 || y2 || y3 || y4;

/*define the module for the mean of the whole vector alpha beta y1-y4*/
/*theta=alpha0,beta0,d11,d12,d22,sigma^2,gamma*/

start meanfunc(theta) global(time,x,p,n);
    alpha0=theta[1];
    beta0=theta[2];
    d11=theta[3];
    d12=theta[4];
    d22=theta[5];
    sigma2=theta[6];
    gamma1=theta[7:(6+p)];

    coeff=0 || 1 || time[1] || time[2] || time[3] || time[4];
    coeff2=1 || 0 || 1 || 1 || 1 || 1;

```

```

    prod=alpha0*coeff2 + beta0 * coeff;;
    sterm=prod;
    do i=2 to n;
        sterm=sterm//prod;
    end;

    temp2=x*gamma1; /*n*1 vector*/
    fterm=j(n,1,0) || j(n,1,0) || temp2 || temp2 ||temp2 ||temp2
    ;

    mean=fterm+sterm;
    return(mean);
finish meanfunc;

start varfunc(theta) global(time, n,p);
    alpha0=theta[1];
    beta0=theta[2];
    d11=theta[3];
    d12=theta[4];
    d22=theta[5];
    sigma2=theta[6];
    gamma=theta[7:(6+p)];

    coeff=(I(2) || j(2,4,0)) // (j(4,1,1) || time[1:4] || I(4));
    variance=j(6,6,0);
    variance[1,1]=d11;
    variance[1,2]=d12;
    variance[2,1]=d12;
    variance[2,2]=d22;
    variance[3:6,3:6]=sigma2*I(4);

    var=coeff*variance*coeff`;
    return(var);
finish varfunc;

```

```

start hfunc(theta) global(time,x,p,n,y1,y2,y3,y4,y5,c1,c2,c3,c4,rindex2,
rindex3,rindex4,y12,y123,y1234);
    alpha0=theta[1];
    beta0=theta[2];
    d11=theta[3];
    d12=theta[4];
    d22=theta[5];
    sigma2=theta[6];
    gamma=theta[7:(6+p)];

    mean=meanfunc(theta); /*n*6 matrix: mean of alpha beta y1-y4*/
    /
    var=varfunc(theta);      /*6*6 matrix*/

    fixed=x*gamma;

    h1=j(n,1,.);
    h2=j(n,1,.);
    h3=j(n,1,.);
    h4=j(n,1,.);

    h1=mean[,1]+(y1-mean[,5])*inv(var[5,5])*var[1,5]`; /*conditional
    mean of y5 given y1*/

    temp1=mean[,1:2] + (y1-mean[,3])*inv(var[3,3])*var[3,1:2];
    h1=fixed+ temp1*(1//time[5]); /*conditional mean of y5 given y1*/
    /

    temp2=j(n,2,.);
    temp2[rindex2,]=mean[rindex2,1:2] + (y12[rindex2,]-mean[rindex2
    ,3:4])*inv(var[3:4,3:4])*var[3:4,1:2];
    h2[rindex2,]=fixed[rindex2,] + temp2[rindex2,]*(1//time[5]);
    temp3=j(n,2,.);
    temp3[rindex3,]=mean[rindex3,1:2] + (y123[rindex3,]-mean[rindex3
    ,3:5])*inv(var[3:5,3:5])*var[3:5,1:2];

```

```

h3[rindex3,]=fixed[rindex3,] + temp3[rindex3,]*(1//time[5]);

temp4=j(n,2,.);
temp4[rindex4,]=mean[rindex4,1:2] + (y1234[rindex4,]-mean[rindex4
,3:6])*inv(var[3:6,3:6])*var[3:6,1:2];
h4[rindex4,]=fixed[rindex4,] + temp4[rindex4,]*(1//time[5]);

h=h1 || h2 || h3 || h4;
* print h;
return(h);
finish hfunc;

start htilda(theta) global (time,x,p,tp,p1,p2,p3,p4,n,y1,y2,y3,y4,y5,x2,
lambda1,lambda2,lambda3,lambda4,k1,k2,k3,k4,a,c1,c2,c3,c4,
x_lam1,x_lam2,x_lam3,x_lam4, rindex2,rindex3,rindex4,y12,y123,y1234);
alpha0=theta[1];
beta0=theta[2];
d11=theta[3];
d12=theta[4];
d22=theta[5];
sigma2=theta[6];
gamma=theta[7:(6+p)];
c=theta[(7+p):tp]; /*c is the theta on page 21: lambda1 lambda2
lambda3 lambda4*/

/*define the design matrix for lambda1, lambda2, lambda3,
lambda4*/
delta1=c[1:p1];
delta2=c[(p1+1):(p1+p2)];
delta3=c[(p1+p2+1):(p1+p2+p3)];
delta4=c[(p1+p2+p3+1):(p1+p2+p3+p4)];

h=hfunc(theta[1:(6+p)]);
h1=h[,1];
h2=h[,2];

```



```

h3=h[,3];
h4=h[,4];

h1_tilda=j(n,1,.);
h2_tilda=j(n,1,.);
h3_tilda=j(n,1,.);
h4_tilda=j(n,1,.);

h1_tilda=h1-(1-lambda1)#(x_lam1*delta1);

h2_tilda[rindex2,]=h2[rindex2,]-(1-lambda2[rindex2,])#k1[
rindex2,]#(x_lam2[rindex2,]*delta2);
h3_tilda[rindex3,]=h3[rindex3,]-(1-lambda3[rindex3,])#k2[
rindex3,]#(x_lam3[rindex3,]*delta3);
h4_tilda[rindex4,]=h4[rindex4,]-(1-lambda4[rindex4,])#k3[rindex4
,]#(x_lam4[rindex4,]*delta4);

h_tilda=h1_tilda || h2_tilda || h3_tilda || h4_tilda;
* print htilda;
return(h_tilda);
finish htilda;

/*using numerical derivatives*/
start hder(theta) global(time,x,p,tp,p1,p2,p3,p4,n,y1,y2,y3,y4,y5,x2,
lambda1,lambda2,lambda3,lambda4,k1,k2,k3,k4,a,c1,c2,c3,c4,
x_lam1,x_lam2,x_lam3,x_lam4, rindex2,rindex3,rindex4,y12,y123,y1234);
alpha0=theta[1];
beta0=theta[2];
d11=theta[3];
d12=theta[4];
d22=theta[5];
sigma2=theta[6];
gamma=theta[7:(6+p)];
c=theta[(7+p):tp]; /*c is the theta on page 21: lambda1 lambda2

```

```

lambda3 lambda4*/

    theta2=theta[1:(6+p)];
h=hfunc(theta2);
    h1=h[,1];
h2=h[,2];
h3=h[,3];
h4=h[,4];

    mat=I(6+p);
    eps=0.000001;
    h1_der=j(n,tp,0);
h2_der=j(n,tp,.);
h3_der=j(n,tp,.);
h4_der=j(n,tp,.);

    do i=1 to (6+p);
        theta_temp=theta2+eps*mat[i,]`;
        hplus=hfunc(theta_temp);
        h1_der[,i]=(hplus[,1]-h1)/eps;
        h2_der[,i]=(hplus[,2]-h2)/eps;
        h3_der[,i]=(hplus[,3]-h3)/eps;
        h4_der[,i]=(hplus[,4]-h4)/eps;
    end;

h1_der[, (7+p):(6+p+p1)]=-(1-lambda1)#x_lam1;
h2_der[rindex2, (7+p):tp]=0;
h3_der[rindex3, (7+p):tp]=0;
h4_der[rindex4, (7+p):tp]=0;

    h2_der[rindex2, (7+p+p1):(6+p+p1+p2)]=-(1-lambda2[rindex2,])#
    k1[rindex2,]#x_lam2[rindex2,];
h3_der[rindex3, (7+p+p1+p2):(6+p+p1+p2+p3)]=-(1-lambda3[rindex3
,])#k2[rindex3,]#x_lam3[rindex3,];

```

```

        h4_der[rindex4, ((7+p+p1+p2+p3)):tp] = -(1-lambda4[rindex4,]) #k3[
        rindex4,] #x_lam4[rindex4,];

    h_der=h1_der//h2_der//h3_der//h4_der;
    *   print h_der;
    return(h_der);
finish hder;

start objfunc(theta) global(time,x,p,tp,p1,p2,p3,p4,n,y1,y2,y3,y4,y5,x2,
lambda1,lambda2,lambda3,lambda4,k1,k2,k3,k4,a,c1,c2,c3,c4,
x_lam1,x_lam2,x_lam3,x_lam4, rindex2,rindex3,rindex4,y12,y123,y1234);
    alpha0=theta[1];
    beta0=theta[2];
    d11=theta[3];
    d12=theta[4];
    d22=theta[5];
    sigma2=theta[6];
    gamma=theta[7:(6+p)];
    c=theta[(7+p):tp]; /*c is the theta on page 21: lambda1
    lambda2 lambda3 lambda4*/

    htilda=htilda(theta);
    h1_tilda=htilda[,1];
    h2_tilda=htilda[,2];
    h3_tilda=htilda[,3];
    h4_tilda=htilda[,4];

    hder=hder(theta);
    h1_der=hder[1:n,];
    h2_der=hder[(n+1):(2*n),];
    h3_der=hder[(2*n+1):(3*n),];
    h4_der=hder[(3*n+1):(4*n),];

    object=j(tp,1,0);
    do i=1 to n;

```

```

if c1[i]=0 then
  do;
    temp1=lambda1[i]*h1_der[i,]\'/k1[i];
    q1=-temp1/k1[i];
    object=object+(1-c1[i])*q1*(h2_tilda[i]-h1_tilda[
      i]);
    end;

if c2[i]=0 then
  do;
    temp2=temp1+lambda2[i]*h2_der[i,]\'/k2[i];
    q2=-temp2/k2[i];
    object=object+(1-c2[i])*q2*(h3_tilda[i]-h2_tilda
      [i]);
    end;

if c3[i]=0 then
  do;
    temp3=temp2+lambda3[i]*h3_der[i,]\'/k3[i];
    q3=-temp3/k3[i];
    object=object+(1-c3[i])*q3*(h4_tilda[i]-h3_tilda
      [i]);
    end;

if c4[i]=0 then
  do;
    temp4=temp3+lambda4[i]*h4_der[i,]\'/k4[i];
    q4=-temp4/k4[i];
    object=object+(1-c4[i])*q4*(y5[i]-h4_tilda[i]);
    end;

end;
obj=object\'*object;
obj=log(obj);

return(obj);

```

```
finish objfunc;
```

```
start objfunc2(theta) global(time,x,p,tp,p1,p2,p3,p4,n,y1,y2,y3,y4,y5,x2
,lambda1,lambda2,lambda3,lambda4,k1,k2,k3,k4,a,c1,c2,c3,c4,
x_lam1,x_lam2,x_lam3,x_lam4, rindex2,rindex3,rindex4,y12,y123,y1234);
    alpha0=theta[1];
    beta0=theta[2];
    d11=theta[3];
    d12=theta[4];
    d22=theta[5];
    sigma2=theta[6];
    gamma=theta[7:(6+p)];
    c=theta[(7+p):tp]; /*c is the theta on page 21: lambda1 lambda2
lambda3 lambda4*/

    htilda=htilda(theta);
    h1_tilda=htilda[,1];
    h2_tilda=htilda[,2];
    h3_tilda=htilda[,3];
    h4_tilda=htilda[,4];

    hder=hder(theta);
    h1_der=hder[1:n,];
    h2_der=hder[(n+1):(2*n),];
    h3_der=hder[(2*n+1):(3*n),];
    h4_der=hder[(3*n+1):(4*n),];

    object=j(tp,1,0);
    do i=1 to n;
        if c1[i]=0 then
            do;
                temp1=lambda1[i]*h1_der[i,]`/k1[i];
                q1=-temp1/k1[i];
                object=object+(1-c1[i])*q1*(h2_tilda[i]-h1_tilda[
i]);
```

```

end;

if c2[i]=0 then
do;
temp2=temp1+lambda2[i]*h2_der[i,]\'/k2[i];
q2=-temp2/k2[i];
object=object+(1-c2[i])*q2*(h3_tilda[i]-h2_tilda
[i]);
end;

if c3[i]=0 then
do;
temp3=temp2+lambda3[i]*h3_der[i,]\'/k3[i];
q3=-temp3/k3[i];
object=object+(1-c3[i])*q3*(h4_tilda[i]-h3_tilda
[i]);
end;

if c4[i]=0 then
do;
temp4=temp3+lambda4[i]*h4_der[i,]\'/k4[i];
q4=-temp4/k4[i];
object=object+(1-c4[i])*q4*(y5[i]-h4_tilda[i]);
end;

* print object;
end;

return(object);
finish objfunc2;

```

```

start obji(theta) global(s,time,x,p,tp,p1,p2,p3,p4,n,y1,y2,y3,y4,y5,x2,
lambda1,lambda2,lambda3,lambda4,k1,k2,k3,k4,a,c1,c2,c3,c4,
x_lam1,x_lam2,x_lam3,x_lam4, rindex2,rindex3,rindex4,y12,y123,y1234,h1_
tilda,h2_tilda,h3_tilda,h4_tilda,h1_der,h2_der,h3_der,h4_der);

```

```

alpha0=theta[1];
beta0=theta[2];
d11=theta[3];
d12=theta[4];
d22=theta[5];
sigma2=theta[6];
gamma=theta[7:(6+p)];
c=theta[(7+p):tp]; /*c is the theta on page 21: lambda1 lambda2
lambda3 lambda4*/

i=s;
  object=j(tp,1,0);
    if c1[i]=0 then
      do;
        temp1=lambda1[i]*h1_der[i,]'/k1[i];
        q1=-temp1/k1[i];
        object=object+(1-c1[i])*q1*(h2_tilda[i]-h1_tilda[
          i]);
        end;

    if c2[i]=0 then
      do;
        temp2=temp1+lambda2[i]*h2_der[i,]'/k2[i];
        q2=-temp2/k2[i];
        object=object+(1-c2[i])*q2*(h3_tilda[i]-h2_tilda[
          i]);
        end;

    if c3[i]=0 then
      do;
        temp3=temp2+lambda3[i]*h3_der[i,]'/k3[i];
        q3=-temp3/k3[i];
        object=object+(1-c3[i])*q3*(h4_tilda[i]-h3_tilda[
          i]);
        end;

```

```

        if c4[i]=0 then
            do;
                temp4=temp3+lambda4[i]*h4_der[i,]`/k4[i];
                q4=-temp4/k4[i];
                object=object+(1-c4[i])*q4*(y5[i]-h4_tilda[i]);
            end;
        *   print object;

        return(object);
finish obji;

        *x0=j(tp,1,0);
        x0=j(tp,1,0.12);

x0[1:(6+p)]={ 0.5309, -0.00887,  0.03424, 0.004190, 0.001460,  0.03023,
0.03508,    0.2306, -0.09488, -0.07592};

print x0;
nrow=nrow(x0);
print nrow;

        optn2=j(1,2,.);  optn2[1]=0; optn2[2]=1;
        tc=j(1,10,.);
        tc[1]=200;
        *   tc[2]=500;
            tc[2]=1000;

call nlpqn(rc,xres,"objfunc",x0,optn2) tc=tc;

/*calculate the estimated mean at the last time point*/
        * C: coarsening variable;
        C=j(n,1,1);
        do i = 1 to n;
            if c1[i]=0 then c[i]=2;
            if c2[i]=0 then c[i]=3;

```



```

        if c3[i]=0 then c[i]=4;
        if c4[i]=0 then c[i]=5;
end;

h=hfunc(xres[1:(6+p)]');
h1=h[,1]; h2=h[,2]; h3=h[,3]; h4=h[,4];

mkh=0;
mk=0;
do i=1 to n;
    temp1=((c[i]=1)-lambda1[i]*(c[i]>=1))/k1[i];
    temp2=((c[i]=2)-lambda2[i]*(c[i]>=2))/k2[i];
    temp3=((c[i]=3)-lambda3[i]*(c[i]>=3))/k3[i];
    temp4=((c[i]=4)-lambda4[i]*(c[i]>=4))/k4[i];
    mkh=mkh+temp1*h1[i]+(temp2*h2[i])*(c[i]>=2)+(temp3*h3[i])*(c[i]>=3)+(
    temp4*h4[i])*(c[i]>=4);
    mk=mk+temp1*(c[i]>=1)+temp2*(c[i]>=2)+temp3*(c[i]>=3)+temp4*(c[i]>
    =4);
end;

numfirst=sum(y5/k4);
testipw=numfirst/n;
print testipw;
denfirst=sum((c=5)/k4);
print denfirst;

print mkh mk;
mu=(numfirst+mkh)/(denfirst+mk);
print mu;

tp1=tp+p1+p2+p3+p4+1;
print tp1;
/*****
    calculate the standard error
*****/

```

```

htilda=htilda(xres`);
h1_tilda=htilda[,1];
h2_tilda=htilda[,2];
h3_tilda=htilda[,3];
h4_tilda=htilda[,4];

hder=hder(xres`);
  h1_der=hder[1:n,];
h2_der=hder[(n+1):(2*n),];
h3_der=hder[(2*n+1):(3*n),];
h4_der=hder[(3*n+1):(4*n),];

B_n=j(tp1,tp1,0);
  do s=1 to n;

    m_temp=j(tp1,1,0);
    m_temp[1:p1]=(c1[s]-lambda1[s])#x_lam1[s,]`;
    if c1[s]=0 then
      m_temp[(p1+1):(p1+p2)]=(c2[s]-lambda2[s])#x_
        lam2[s,]`;
    if c2[s]=0 then
      m_temp[(p1+p2+1):(p1+p2+p3)]=(c3[s]-lambda3[s])#x_lam3[s
        ,]`;
      if c3[s]=0 then
        m_temp[(p1+p2+p3+1):(p1+p2+p3+p4)]=(c4[s]-
          lambda4[s])#x_lam4[s,]`;

    m_temp[(p1+p2+p3+p4+1):(tp1-1)]=obji(xres`);

    temp1=((c[s]=1)-lambda1[s]*(c[s]>=1))/k1[s];
    temp2=((c[s]=2)-lambda2[s]*(c[s]>=2))/k2[s];
    temp3=((c[s]=3)-lambda3[s]*(c[s]>=3))/k3[s];
    temp4=((c[s]=4)-lambda4[s]*(c[s]>=4))/k4[s];
    mkh=(c[s]>=1)*(temp1*(h1[s]-mu))+(c[s]>=2)*(temp2*(h2[s]-

```

```

mu))+(c[s]>=3)*(temp3*(h3[s]-mu))+(c[s]>=4)*(temp4*(h4[s]-mu));
m_temp[tp1]=(c[s]=5)*((y5[s]-mu)/k4[s])+m_kh;

B_n=B_n+m_temp*m_temp`;
end;
B_n=B_n/n;

/*Then we compute the matrix A_n by taking the derivatives of
the m-estimating equations. */

eps=1e-6;
mat=I(tp);
der=hder[,1:(6+p)];

A_n=j(tp1,tp1,0);
do s=1 to n;
    a_temp=j(tp1,tp1,0);
    a_temp[1:p1,1:p1]=lambda1[s]#(1-lambda1[s])#x_lam1[s,]*x_lam1[s,];
    if c1[s]=0 then
        a_temp[(p1+1):(p1+p2),(p1+1):(p1+p2)]=lambda2[s]#(1-lambda2[s])#x_lam2[s,]*x_lam2[s,];
    if c2[s]=0 then
        a_temp[(p1+p2+1):(p1+p2+p3),(p1+p2+1):(p1+p2+p3)]=lambda3[s]#(1-lambda3[s])#x_lam3[s,]*x_lam3[s,];
    if c3[s]=0 then
        a_temp[(p1+p2+p3+1):(p1+p2+p3+p4),(p1+p2+p3+1):(p1+p2+p3+p4)]=lambda4[s]#(1-lambda4[s])#x_lam4[s,]*x_lam4[s,];

    mtemp=obji(xres`);
    do u=1 to tp;
        x0temp=xres`+eps*mat[,u];
        mtempplus=obji(x0temp);
        a_temp[(u+p1+p2+p3+p4),(p1+p2+p3+p4+1):(tp1-1)]=- (mtempplus`-mtemp`)/eps;

```

```

end;

fterm1=j(1,p1,0);
fterm2=j(1,p2,0);
fterm3=j(1,p3,0);
fterm4=j(1,p4,0);

    if c[s]=5 then do;
        fterm1=-(c[s]=5)*(y5[s]-mu)*lambda1[s]/k4[
            s]*x_lam1[s,];
        fterm2=-(c[s]=5)*(y5[s]-mu)*lambda2[s]/k4[s]*x_lam2[s
            ,];
        fterm3=-(c[s]=5)*(y5[s]-mu)*lambda3[s]/k4[s]*x_lam3[s
            ,];
        fterm4=-(c[s]=5)*(y5[s]-mu)*lambda4[s]/k4[s]*x_lam4[s
            ,];

    end;

dm1=((c[s]=1)-lambda1[s]*(c[s]>=1))*(c[s]>=1);
dm2=((c[s]=2)-lambda2[s]*(c[s]>=2))*(c[s]>=2);
dm3=((c[s]=3)-lambda3[s]*(c[s]>=3))*(c[s]>=3);
dm4=((c[s]=4)-lambda4[s]*(c[s]>=4))*(c[s]>=4);

    a_temp[tp1,1:p1]=fterm1;
    if c[s]>=1 then
        a_temp[tp1,1:p1]=a_temp[tp1,1:p1] + ((
            c[s]>=1)*(1-lambda1[s])-dm1)*lambda1[s
                ]*(h1[s]-mu)*x_lam1[s,]/k1[s];
    if c[s]>=2 then
        a_temp[tp1,1:p1]=a_temp[tp1,1:p1] -
            dm2*lambda1[s]*(h2[s]-mu)*x_lam1[s,]/
            k2[s];
    if c[s]>=3 then
        a_temp[tp1,1:p1]=a_temp[tp1,1:p1] -
            dm3*lambda1[s]*(h3[s]-mu)*x_lam1[s,]/

```

```

k3[s];
if c[s]>=4 then
a_temp[tp1,1:p1]=a_temp[tp1,1:p1] -
dm4*lambda1[s]*(h4[s]-mu)*x_lam1[s,]/
k4[s];

a_temp[tp1,(p1+1):(p1+p2)]=fterm2;
if c[s]>=2 then
a_temp[tp1,(p1+1):(p1+p2)]=a_temp[tp1,(p1+1):(p1+p2)
] + ((c[s]>=2)*(1-lambda2[s])-dm2)*lambda2[s]*(h2[s
]-mu)*x_lam2[s,]/k2[s];
if c[s]>=3 then
a_temp[tp1,(p1+1):(p1+p2)]=a_temp[tp1,(p1+1):(p1+p2)
] - dm3*lambda2[s]*(h3[s]-mu)*x_lam2[s,]/k3[s];
if c[s]>=4 then
a_temp[tp1,(p1+1):(p1+p2)]=a_temp[tp1,(p1+1):(p1+p2)
] - dm4*lambda2[s]*(h4[s]-mu)*x_lam2[s,]/k4[s];

a_temp[tp1,(p1+p2+1):(p1+p2+p3)]=fterm3;
if c[s]>=3 then
a_temp[tp1,(p1+p2+1):(p1+p2+p3)]=a_temp[tp1,(p1+p2
+1):(p1+p2+p3)] + ((c[s]>=3)*(1-lambda3[s])-dm3)*
lambda3[s]*(h3[s]-mu)*x_lam3[s,]/k3[s];
if c[s]>=4 then
a_temp[tp1,(p1+p2+1):(p1+p2+p3)]=a_temp[tp1,(p1+p2
+1):(p1+p2+p3)] - dm4*lambda3[s]*(h4[s]-mu)*x_lam3
[s,]/k4[s];

a_temp[tp1,(p1+p2+p3+1):(p1+p2+p3+p4)]=fterm4;
if c[s]>=4 then
a_temp[tp1,(p1+p2+p3+1):(p1+p2+p3+p4)]=a_temp[
tp1,(p1+p2+p3+1):(p1+p2+p3+p4)] + ((c[s]>=4)*(1-

```

```

lambda4[s]) - dm4) * lambda4[s] * (h4[s] - mu) * x_lam4[s
,] / k4[s];

hder_1 = der[s,];
hder_2 = der[(s+n),];
hder_3 = der[(s+2*n),];
hder_4 = der[(s+3*n),];

k1s = (c[s] >= 1) * (dm1 * hder_1 / k1[s]);
k2s = (c[s] >= 2) * (dm2 * hder_2 / k2[s]);
k3s = (c[s] >= 3) * (dm3 * hder_3 / k3[s]);
k4s = (c[s] >= 4) * (dm4 * hder_4 / k4[s]);
a_temp[tp1, (p1+p2+p3+p4+1):tp] = -k1s - k2s - k3s - k4s;

a_temp[tp1, tp1] = 1;

A_n = A_n + a_temp;
end;

A_n = A_n / n;
A_inv = ginv(A_n);
V_n = A_inv * B_n * A_inv' / n;
sderr = sqrt(V_n[tp1, tp1]);

print sderr;

quit;

```