ABSTRACT

BRINKLEY, JASON S. Generalized Estimators of the Attributable Benefit of an Optimal Treatment Regime. (Under the direction of Professor Anastasios Tsiatis).

For many diseases where there are several treatment options it is often the case that there is no consensus on the best treatment to give to individual patients. In such cases it may be necessary to define a strategy for treatment assignment; that is, an algorithm which dictates the treatment an individual should receive based on their measured characteristics. Such a strategy or algorithm is also referred to as a treatment regime. The optimal treatment regime is the strategy that would provide the most public health benefit by minimizing as many poor outcomes as possible. Using a measure that is a generalization of attributable risk and notions of potential outcomes, we derive two estimators for the proportion of events that could have been prevented had the optimal treatment regime been implemented. Traditional attributable risk studies look at the added risk that can be attributed to exposure of some contaminant, here we will instead study the benefit that can be attributed to using the optimal treatment strategy.

We will show how regression models can be used to estimate the optimal treatment strategy and the attributable benefit of that strategy. While in some cases this is a very good method of estimation, it can be unstable if the regression model is not correctly specified. In trying to reduce the potential for bias a doubly robust estimator will be introduced that offers some protection from model misspecification. We derive the large sample properties of each estimator and explore the pros and cons of each via simulation studies.

As a motivating example we will apply our methods to an observational study of 3856 patients treated at the Duke University Medical Center with prior coronary artery bypass graft surgery (CABG) and presented with further heart related problems requiring a catheterization. The patients may be treated with either medical therapy alone (MED) or a combination of medical therapy and percutaneous coronary intervention (PCI) without general consensus on which is the best treatment for individual patients.

Generalized Estimators of the Attributable Benefit of an Optimal Treatment Regime

by
Jason S. Brinkley

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

Statistics

Raleigh, North Carolina

2008

APPROVED BY:

Dr. Wenbin Lu	Dr. Daowen Zhang
Dr. Anastasios Tsiatis Chair of Advisory Committee	Dr. Marie Davidian

DEDICATION

For Sandra Jeanne

BIOGRAPHY

From the first time Santa left a Rubik's Cube in his stocking, Jason Brinkley has been obsessed with solving puzzles. His love of numbers led him to East Carolina University, where he received a B.A. in mathematics in 2001. Continuing his studies at ECU, Brinkley was awarded a M.A. in mathematics in 2003. He enrolled at N.C. State that fall, and was awarded a M.S. in Statistics in 2005, and is set to receive his doctorate in the fall of 2008. When he's not working a Sudoku, Jason is spending time with his wife and 18-month-old son, trying to get them both to say binomial coefficient.

ACKNOWLEDGMENTS

This body of work would not have been completed without the help of a lot of people. First and foremost I want to thank my advisor, Dr. Butch Tsiatis; his insight and innovation were the motivation for this work. I would also like to thank Dr. Kevin Anstrom from the Duke Clinical Research Institute for helping me get access to real-life data with questions of clinical interest that could be addressed by the methods presented here. Kevin's help in coming up with appropriate simulation studies and interpreting data analysis was invaluable. My hope is to continue to work with both Butch and Kevin on ways to expand the methods discussed here so that they will aid the medical community in assessing the benefits of treatment.

Special thanks to Dr. Pam Arroway and Dr. Bill Swallow for all of their counsel during my graduate studies at NCSU, and Dr Katalin Szucs of East Carolina University for encouraging me to pursue my Ph.D. I would like to thank my fellow graduate students at NCSU for all their help, especially Laine Elliot and Liz Nelson. Laine and Liz always made time to discuss ideas, help in the organization and technical issues of drafting the thesis, and being available to occasionally baby-sit so I could do some work. Finally, I would like to thank my wonderful wife and family for all of their encouragement; their unwavering faith in me continues to give me strength.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Introduction

Since its inception by Levin [6], attributable risk (AR) has become a widely used measure of association between risk factors and disease status. Jewell [5] denotes the population attributable risk as the fraction of all disease cases that can be linked to exposure from some factor. AR has been referred to by many other names (e.g. 'etiologic fraction' or 'attributable fraction') and has a long history within the medical and epidemiology communities.

Observational data can be used to assess the impact of various treatment strategies on patient health outcomes, where a treatment strategy is defined as an algorithm which dictates the treatment an individual should receive based on their characteristics. Furthermore, we define the optimal treatment strategy to be the strategy that will result in the smallest proportion of poor outcomes (events) among the study population. In this thesis we propose a generalization of the attributable risk measure which is defined as the fraction of observed events that could have been prevented had the optimal treatment strategy been implemented. In contrast to traditional AR studies where the goal is to measure the added risk attributed to exposure of some contaminant, this measure is intended to study the positive public health impact of a treatment strategy. As such we will refer to it as attributable benefit (AB).

1.1 Example

Our motivating problem comes from the Duke Databank for Cardiovascular Diseases (DDCD), a large database housed at the Duke University Medical Center (DUMC).

Started in 1969, the DDCD collects data on the in-hospital clinical course of all patients who undergo cardiac catheterization, an interventional cardiac procedure or coronary artery bypass surgery at Duke University Medical Center. A wide variety of baseline patient information is collected and the DDCD attempts to collect follow-up data on patients with at least one diseased vessel six months and one year after hospitalization and yearly thereafter. Data are collected in the form of a mailed questionnaire or surveyed by phone for those who do not respond to the mailing [1]. One of the largest and oldest such databases, the DDCD provides essential information to cardiologists about the effectiveness of various treatments on patient health.

In this manuscript, we are focused on the subset of patients with a previous coronary artery bypass graft (CABG) surgery and requiring a later catheterization due to continued symptoms. For technical reasons, the vast majority of patients are not considered to be candidates for a second CABG surgery and therefore their treatment options are limited to either medical therapies (MED) or some combination of medical therapies and percutaneous coronary intervention (PCI) which is also known as angioplasty. A question of clinical interest is whether PCI provides some benefit (here in the form of lower 1-year mortality rates) above optimal medical therapy for either the overall study population or some subset defined by clinical characteristics.

In studying this question we will explore the effects that covariates such as patient demographics (i.e., age and gender), health status (i.e., body mass index, pulse, patient and family history of coronary artery disease), and cardiovascular measurements (i.e., ejection fraction, number of diseased vessels) play as potential confounders. To see the benefit of using our estimator we will compare estimators from models with zero, one, and several confounding variables. The goal of which is to explore the impact of an optimal treatment strategy while simultaneously examining the effects of confounding. The benefit of the optimal treatment regime will be estimated by the proportion of one-year deaths that could have been prevented if that optimal regime had been followed. The study population includes 3856 individuals with a follow-up catheterization between the years 1986-2001 and a history of CABG surgery.

1.2 Methods

For simplicity we consider only the case where there are two treatment options which will be labeled by the treatment indicator $E = 0$ or 1. We denote the event of interest by the binary indicator D where $D = 1$ denotes poor outcome and $D = 0$ denotes good outcome. Let X denote the vector of covariates measured on each individual that may be used to determine which treatment they receive, we denote a treatment strategy g as a function that maps values of X to 0 or 1. That is, $g(x)$ denotes the treatment E (either 1 or 0) that will be assigned to an individual with covariates $X = x$.

The goals of this thesis are to define and estimate the optimal treatment strategy $g_{opt}(x)$; i.e., the treatment strategy which results in the smallest probability of a poor outcome, and to estimate the attributable benefit of using the optimal regime. We will show that our proposed measure is a generalization of the standard attributable risk. We will formally define the optimal treatment strategy and develop an estimator for attributable benefit that is a function of an individuals covariates. In order to formally define the parameter of interest and the required assumptions, we will use the ideas of potential outcomes and causal inference as described by Neyman [7] and Rubin [8].

In studying the attributable benefit we will define two different estimators. The first estimator will be based on utilizing an appropriate statistical model, while the second estimator will utilize two different statistical models in a way to try and reduce bias. We will examine the large sample properties of these estimators and two strategies for developing confidence intervals are derived. The behavior of the proposed estimators are described in simulation studies and each estimator will be applied to a real life example. Later we intend to compare the merits and limitations of each estimator. We conclude with the limitations of the current methods and possible avenues of future research.

The remainder of this chapter is devoted to introducing the concepts and notations that will be important in discussing AB and its estimators. As such we will need to provide overviews of attributable risk, statistical causality, optimal treatment regimes, m-estimation and influence functions of m-estimators. These concepts are used extensively throughout the thesis and some attention needs to be given to the ideas and notation used in each.

1.3 Potential Outcomes and Statistical Causality

Using the ideas and notation developed by Rubin [8] we define the potential outcomes $D^*(1)$ and $D^*(0)$ to denote the indicator of a poor outcome (event) for an arbitrary individual in our population had they received treatment 1 or 0 (respectively). In actuality, at most one of these potential outcomes can be observed for any given individual. In addition, we also collect baseline covariates X on these individuals. Consequently, the potential outcomes for any individual will be defined as $W = \{D^*(1), D^*(0), X\}$ and it is assumed that there is some underlying population distribution for these potential outcomes.

In contrast, the observable data will be denoted by $O = (D, E, X)$ where E denotes the treatment (0 or 1) that was actually given to an individual, D denotes the outcome indicator that resulted from that treatment, and X is a single or vector of baseline covariates measured on the individuals. The key assumptions that will allow us to derive important aspects of the distribution of the potential outcomes, which are not directly observable, from the distribution of the observable data were proposed by Rubin [8].

The first assumption, also known as the Stable Unit Treatment Value Assumption (SUTVA), states that the observed outcome for an individual is the same as the potential outcome for the corresponding treatment that the individual received. In terms of the current framework, the SUTVA assumption can be written as $D = D^*(1)E + D^*(0)(1-E)$. The second assumption is also referred to as the strong ignorability assumption or the no unmeasured confounders assumption. This assumption states that the treatment E assigned to an individual is independent of the potential outcomes $\{D^*(1), D^*(0)\}$ conditional on the covariates X. In an observational study the treatment that a physician chooses to give a patient can be reasonably assumed to only be based on characteristics of the patient at the time of treatment and not on the patient's potential outcome (which of course is not known at the time of treatment). Consequently, this second assumption will be tenable if the key factors influencing the decision making process for a physician was captured in the data X that were collected. If, however, there are additional factors beyond the data X that influence treatment decisions, then this assumption may not hold. For the rest of this paper we will assume both SUTVA and strong ignorability do indeed hold.

Under these assumptions, for example, we can compute the marginal probability

$P\{D^*(0) = 1\}$ from the distribution of the observed data as

$$\begin{aligned} P\{D^*(0) = 1\} &= E_X\{P[D^*(0) = 1|X]\} \\ &= E_X\{P[D^*(0) = 1|X, E = 0]\} \\ &= E_X\{P(D = 1|X, E = 0)\}, \end{aligned}$$

where we use the notation $E_X(\cdot)$ to emphasize that this expectation is taken with respect to the marginal distribution of X. Similarly, we can show that

$$P\{D^*(1) = 1\} = E_X\{P(D = 1|X, E = 1)\}.$$

1.4 Attributable Risk

As stated earlier, attributable risk has long been a useful measure in the study of the interaction between exposure and disease. Sometimes referred to as the population attributable fraction or population etiologic fraction [2]; its this co-mingling of different names and concepts that has led to a lot of confusion over using AR in practice. We will consider attributable risk to be the proportion of disease cases that can be attributed to exposure ($E = 1$), which can be written as

$$AR = \frac{P(D = 1) - P(D = 1|E = 0)}{P(D = 1)}. \tag{1.1}$$

For most practical applications, the above estimate of AR is biased because it does not take into account confounders that may affect both exposure and disease status. That is to say there may be some variable, X, that plays an important role in someone's exposure and disease status. Therefore the real quantity of public health interest is a version of AR that accounts for confounders. One popular way to write AR that is adjusted for with multiple levels (say $X = x_1, ...x_C$) is from Whittemore [12]

$$AR = \frac{P(D = 1) - \sum_{c=1}^{C} P(X = x_c)P(D = 1|E = 0, X = x_c)}{P(D = 1)} \tag{1.2}$$

which we will write as

$$AR = \frac{P(D = 1) - E_X\{P(D = 1|E = 0, X)\}}{P(D = 1)}, \tag{1.3}$$

where the expectation is with respect to the variable X. While (1.2) is a popular way of writing AR, note that the representation in (1.3) is more useful for dealing with continuous covariates or an entire vector of confounders. AR has proven to be most popular in case-control studies where the measure is most interpretable. Many methods for estimating adjusted AR stem from use in such a design, such as taking weighted sums or rewriting AR in terms of the relative risk measure and estimating that instead. The interested reader may see Benichou [2] for a complete review of adjusted estimators for AR.

We will instead think of adjusted AR as given in (1.3) because we can apply the SUTVA and no unmeasured confounders assumptions of statistical causality to write adjusted AR as

$$AR = \frac{P(D=1) - P\{D^*(0)=1\}}{P(D=1)}. \tag{1.4}$$

In forming variance estimates and confidence intervals a wide range of methods and transformations have been introduced. Recently Graubard and Fears [4] used influence function theory to compute large sample variance estimates based on an estimate of AR that uses statistical models. We will employ similar techniques in this work to show our estimates are asymptotically normal with very few assumptions. Walter [2] suggested to form intervals based on estimates of $\log(1-AR) = \log[P\{D^*(0)=1\}/P(D=1)]$ for better coverage.

1.5 Optimal Treatment Regime

We formally defined a treatment regime to be a function g which maps the values of X to 0 or 1. Hence the treatment regime would assign treatment $g(x)$ (either 0 or 1) to an individual whose covariates $X = x$. We also denote by $\mathcal{G}$ the class of all such treatment regimes. We can now define the potential outcome for an arbitrary individual in our population under the hypothetical situation that treatment regime g was being implemented by $D^*\{g(X)\} = D^*(1)\mathrm{I}\{g(X)=1\} + D^*(0)\mathrm{I}\{g(X)=0\}$ where $\mathrm{I}(\cdot)$ denotes the indicator function. The most simple strategies would be g_0 or g_1 where all patients were given treatment 0 or 1 (respectively) regardless of the values of their covariates X.

We now define the optimal strategy as the one which would result in the smallest probability of a poor outcome. That is,

$$g_{opt}(X) = \arg\min_{g\in\mathcal{G}} \left(E[D^*\{g(X)\}]\right).$$

Under the SUTVA and no unmeasured confounders assumption, it is straightforward to show that

$$\begin{aligned} &E[D^*\{g(X)\}] = P\{D^*\{g(X)\} = 1\} \\ &= E_X[P(D=1|E=1,X)\mathrm{I}\{g(X)=1\} + P(D=1|E=0,X)\mathrm{I}\{g(X)=0\}], \quad (1.5) \end{aligned}$$

and the optimal treatment regime is given by

$$g_{opt}(X) = \mathrm{I}\{P(D=1|E=1,X) - P(D=1|E=0,X) \leq 0\}. \quad (1.6)$$

That is, the optimal treatment regime would assign treatment 1 to an individual with baseline covariates X if the conditional probability that $D = 1$ given $E = 1$ and X is smaller than the conditional probability $D = 1$ given $E = 0$ and X; otherwise give treatment 0.

Now that we have some notation and insight into the optimal treatment regime, we can formally define the parameter of interest. Recall the question of interest was how to estimate the proportion of events that would have been prevented had the optimal strategy been used? We define this quantity as attributable benefit of using the optimal treatment regime. From the notation above we define AB_{opt} as:

$$\begin{aligned} AB_{opt} &= \frac{P(D=1) - P\{D^*(g_{opt})=1\}}{P(D=1)} \quad (1.7) \\ &= 1 - \frac{P\{D^*(g_{opt})=1\}}{P(D=1)}. \quad (1.8) \end{aligned}$$

Since $P(D = 1)$ denotes the overall probability of failure in the population using whatever is the current clinical practice, then AB_{opt} denotes the proportion of these failures that could have been prevented if treatment strategy g_{opt} were used. Because $P(D = 1)$ can be estimated by the simple sample proportion $\bar{D}_n = \sum_{i=1}^{n} D_i/n$ without making any additional assumptions, we will focus attention on estimating the optimal treatment g_{opt} and the proportion of events for that treatment strategy $P\{D^*(g_{opt}) = 1\}$ and then show how these results can be used to derive the estimator and large sample properties for AB_{opt}.

Now consider how to estimate $P\{D^*(g_{opt}) = 1\}$ from the sample of observed data $(D_i, E_i, X_i),\ i = 1, \ldots, n$ assumed to be independent and identically distributed (i.i.d.). Letting $\mathrm{A}(X) = \mathrm{I}\{P(D = 1|E = 1, X) - P(D = 1|E = 0, X) \leq 0\}$, then $D^*(g_{opt}) = D^*(1)\mathrm{A}(X) + D^*(0)\{1 - \mathrm{A}(X)\}$, and by using equations (1.5) and (1.6), we obtain

$$P\{D^*(g_{opt}) = 1\} = E_X\Big[P(D=1|E=1,X)\mathrm{A}(X) + P(D=1|E=0,X)\{1-\mathrm{A}(X)\}\Big]. \quad (1.9)$$

1.6 m-estimation and Influence Functions

Let $Z_1, ..., Z_n$ be identically independently distributed (i.i.d.) random vectors with density $p(z, \beta)$; say we want to estimate β, a k-dimensional parameter. We define $m(Z, \beta)$ as a $k \times 1$ array that satisfy the following properties:

1. $E_\beta\{m(Z, \beta)\} = 0$

2. $E_\beta\{m^T(Z, \beta)m(Z, \beta)\} < \infty$

3. $E_\beta\{m(Z, \beta)m^T(Z, \beta)\}$ is positive definite for all β

where the expectations are taken with respect to the unknown β. We define the m-estimator $\hat{\beta}_n$ of the true value of β (call it β_0) as the solution of the equation $\sum_{i=1}^{n} m(z_i, \hat{\beta}_n) = 0$.

One method for finding the asymptotic variance of an m-estimator is through the use of influence functions. Specifically an estimator $\hat{\beta}_n$ of a vector value parameter β is asymptotically linear if

$$n^{1/2}(\hat{\beta}_n - \beta_0) = n^{-1/2} \sum_{i=1}^{n} \psi_{\hat{\beta}_n}(Z_i) + o_p(1), \tag{1.10}$$

where $\psi_{\hat{\beta}_n}(Z_i)$ is a mean zero random vector with finite variance matrix and $o_p(1)$ are terms that converge in probability to zero as n goes to infinity. The random vector $\psi_{\hat{\beta}_n}(Z_i)$ is referred to as the influence function of the estimator $\hat{\beta}_n$. The representation given in equation (1.10) immediately implies that the estimator is asymptotically normal; that is, $n^{1/2}(\hat{\beta}_n - \beta_0) \to N\left(0, E(\psi\psi^T)\right)$. Furthermore the asymptotic variance can be estimated consistently using the sandwich estimator $\hat{E}(\psi\psi^T) = n^{-1} \sum_{i=1}^{n} \psi_{\hat{\beta}_n}(Z_i)\psi_{\hat{\beta}_n}(Z_i)^T$.

For this work we will be finding influence functions of transformations of certain variables, and as such we should identify some common rules for influence functions of transformations of variables. By a simple application of the delta method, when estimating $\log(\alpha)$, then $\psi_{\log(\hat{\alpha})}(Z_i) = \psi_{\hat{\alpha}}(Z_i)/\alpha$, and when estimating $\alpha_1 - \alpha_2$, $\psi_{\hat{\alpha}_1 - \hat{\alpha}_2}(Z_i) = \psi_{\hat{\alpha}_1}(Z_i) - \psi_{\hat{\alpha}_2}(Z_i)$. The reader may see Tsiatis [11] for a complete overview of the theory of influence functions. Here we will determine the influence function for estimators of several parameters and show how the estimated influence functions will be used to derive the sandwich estimator for the asymptotic variance of our attributable benefit estimators.

1.6.1 Maximum Likelihood Estimators

Generally speaking, maximum likelihood estimators (MLE) and functions of MLE's are indeed m-estimators for their expectations (i.e. $\hat{\beta}_n$ is an m-estimator for $E\{m(Z,\hat{\beta}_n)\} = m(Z,\beta_0)$ where $\hat{\beta}_n$ is the MLE of β). The estimators in this thesis will all be based on functions of maximum likelihood estimators and as such are m-estimators.

Under suitable regularity conditions, the influence function of a m-estimator has a particular form, note that

$$0 = \sum_{i=1}^{n} m(Z_i,\hat{\beta}_n) = \sum_{i=1}^{n} m(Z_i,\beta_0) + \sum_{i=1}^{n} \frac{\partial m(Z_i,\beta_n^*)}{\partial \beta^T}(\hat{\beta}_n - \beta_0)$$

where β_n^* is between $\hat{\beta}_n$ and β_0.

Now note that by a Taylor series expansion we have

$$\begin{aligned} -n^{-1/2}\sum_{i=1}^{n} m(Z_i,\beta_0) &= n^{-1/2}\sum_{i=1}^{n} \frac{\partial m(Z_i,\beta_n^*)}{\partial \beta^T}(\hat{\beta}_n - \beta_0) \\ &= \left\{ n^{-1}\sum_{i=1}^{n} \frac{\partial m(Z_i,\beta_n^*)}{\partial \beta^T} \right\} \left\{ n^{1/2}(\hat{\beta}_n - \beta_0) \right\}. \end{aligned}$$

By Slutsky we can re-write

$$\begin{aligned} n^{1/2}(\hat{\beta}_n - \beta_0) &= \left\{ n^{-1}\sum_{i=1}^{n} \frac{\partial m(Z_i,\beta_n^*)}{\partial \beta^T} \right\}^{-1} \left\{ n^{-1/2}\sum_{i=1}^{n} m(Z_i,\beta_0) \right\} \\ &= n^{-1/2}\sum_{i=1}^{n} \left[-E\left\{ \frac{\partial m(Z_i,\beta_0)}{\partial \beta^T} \right\} \right]^{-1} m(Z_i,\beta_0) + o_p(1). \end{aligned}$$

Thus the influence function is $\psi_{\hat{\beta}_n}(Z_i) = \left[-E\left\{\partial m(Z_i,\beta_0)/\partial\beta^T\right\}\right]^{-1} m(Z_i,\beta_0)$. Consider this under the maximum likelihood framework; specifically the score function $S_\beta(Z_i,\beta) = \partial \log p(Z_i,\beta)/\partial\beta^T$. Then $E_\beta\{S(Z_i,\beta)\} = 0$ and $\sum_{i=1}^{n} S(Z_i,\hat{\beta}_n) = 0$ so the influence function for a maximum likelihood estimator $\hat{\beta}_n$ is

$$\psi_{\hat{\beta}_n}(Z_i) = E\left\{ -\frac{\partial S(Z_i,\beta_o)}{\partial \beta^T} \right\}^{-1} S(Z_i,\beta_0)$$

Applying this to our current situation if $O_i = (D_i, E_i, X_i)$ with $D_i, E_i = 0, 1$; if

$$p(D_i,E_i,X_i,\beta) = \{\mu(X_i,E_i,\beta)\}^{D_i}\{1-\mu(X_i,E_i,\beta)\}^{1-D_i}$$

then the influence function for $\hat{\beta}_n$ is

$$\psi_{\hat{\beta}_n}(O_i) = \left(E\left[\frac{\{\partial \mu_i(\beta)/\partial\beta\}\{\partial \mu_i(\beta)/\partial\beta^T\}}{\mu_i(\beta)\{1-\mu_i(\beta)\}} \right] \right)^{-1} \left[\frac{(\partial \mu_i(\beta)/\partial\beta)}{\mu_i(\beta)\{1-\mu_i(\beta)\}} \right] \{D_i - \mu_i(\beta)\}. \quad (1.11)$$

1.6.2 Multiple m-estimators

Now that we have an idea of how influence functions are generated for a single m-estimator, let us extend this idea to a case involving a vector of m-estimators. The arguments in the previous section extended to cases involving a vector valued β. In some cases it may be easier to partition the parameter of interest (β) into a set of parameters $(\beta_1, ...\beta_k)$, where each β_j may also be vector valued but has a m-estimator that arises naturally. That is to say that if we let $m(Z, \beta_1, ..., \beta_k) = [m_1, ..., m_k]^T$ be a vector of m-estimators for a vector of unknowns $(\beta_1, ..., \beta_k)$. Then it is important to note that while each m_j is the m-estimator for the parameter β_j, it may or may not be a function of multiple β's, that is we may have $m_j = m_j(\beta_1, ..., \beta_k)$. But we will assume that each m_j has the properties aforementioned for m-estimators. In this case $\psi_{\hat{\beta}}$ would be a vector of influence functions for each of the unknown parameters, and we can denote

$$\begin{aligned} \psi_{\hat{\beta}}(Z_i) &= [\psi_{\hat{\beta}_1}(Z_i), ..., \psi_{\hat{\beta}_k}(Z_i)]^T \\ &= \left\{ E\left[\frac{\partial m(Z, \beta_1, ..., \beta_k)}{\partial \beta^T} \right] \right\}^{-1} m(Z, \beta_1, ..., \beta_k) \\ &= \left\{ E\left[\frac{\partial m(Z, \beta_1, ..., \beta_k)}{\partial \beta_1 \partial \beta_2 ... \partial \beta_k} \right] \right\}^{-1} m(Z, \beta_1, ..., \beta_k) \end{aligned}$$

Where $[\partial m/\partial\beta_1\partial\beta_2...\partial\beta_k]$ is the matrix of partial derivatives expressed as

$$\left[\frac{\partial m}{\partial \beta_1 \partial \beta_2 ... \partial \beta_k} \right] = \begin{bmatrix} \partial m_1/\partial\beta_1 & \partial m_1/\partial\beta_2 & ... & \partial m_1/\partial\beta_k \\ \partial m_2/\partial\beta_1 & \partial m_2/\partial\beta_2 & ... & \partial m_2/\partial\beta_k \\ ... & & & \\ \partial m_k/\partial\beta_1 & \partial m_k/\partial\beta_2 & ... & \partial m_k/\partial\beta_k \end{bmatrix}.$$

From this representation if one needs the influence function for the j^{th} β parameter for variance estimates and confidence intervals, then just take the j^{th} row of $\psi_{\hat{\beta}}(Z_i)$.

Chapter 2

Regression Based Estimator

2.1 Introduction

In estimating AB_{opt} it is now clear that we must first estimate $P(D = 1|E, X)$. We will do so by positing a statistical model where

$$P(D = 1|E, X) = \mu(E, X, \beta). \tag{2.1}$$

For simplicity, we will denote $\mu_i(\beta) = \mu(E_i, X_i, \beta)$, $\mu_i(0, \beta) = \mu(0, X_i, \beta)$, and $\mu_i(1, \beta) = \mu(1, X_i, \beta)$. Because D is a binary variable, a natural choice is to use logistic regression models which allow us to study the impact of treatment, covariates, and their interaction. For model (2.1) the parameter β can be estimated using maximum likelihood; that is, the maximum likelihood estimator $\hat{\beta}_n$ would be obtained by maximizing (in β) $\prod_{i=1}^{n}\{\mu_i(\beta)\}^{D_i}\{1-\mu_i(\beta)\}^{1-D_i}$, thus we will denote $\hat{\mu}_i = \mu(E_i, X_i, \hat{\beta}_n)$, $\hat{\mu}_i(1) = \mu(1, X_i, \hat{\beta}_n)$, and $\hat{\mu}_i(0) = \mu(0, X_i, \hat{\beta}_n)$.

As a consequence, a natural estimator for $P\{D^*(g_{opt}) = 1\}$ would be $\hat{P}_n\{D^*(g_{opt}) = 1\} = \hat{E}_n\{D^*(g_{opt}), \hat{\beta}_n\} = n^{-1}\sum_{i=1}^{n}[\hat{\mu}_i(1)\mathrm{A}(X_i, \hat{\beta}_n)+\hat{\mu}_i(0)\{1-\mathrm{A}(X_i, \hat{\beta}_n)\}]$. Where $\mathrm{A}(X_i, \hat{\beta}_n) = \mathrm{I}\{\hat{\mu}_i(1) \leq \hat{\mu}_i(0)\}$, from this we can estimate AB_{opt} with

$$\hat{AB}_{opt} = 1 - [\hat{P}_n\{D^*(g_{opt}) = 1\}/\bar{D}_n]. \tag{2.2}$$

2.2 Large Sample Properties and Standard Errors

In this section we will demonstrate that $\hat{AB}_{opt}$ is asymptotically normal and derive an estimator for its asymptotic variance by finding its influence function. As stated earlier, in order to find the estimated influence function for $\hat{AB}_{opt}$, it suffices to find the estimated influence function of $\log[\hat{E}_n\{D^*(g_{opt}), \hat{\beta}_n\}] - \log(\bar{D}_n)$. For the simple sample average estimator $\bar{D}_n$ of $E(D)$, the influence function is $D_i - E(D)$, the estimated influence function is $D_i - \bar{D}_n$, and the estimated influence function of $\log[\bar{D}_n]$ is $(D_i - \bar{D}_n)/\bar{D}_n$. Therefore deriving the influence function for $\hat{AB}_{opt}$ amounts to finding the influence function for $\hat{E}_n\{D^*(g_{opt}), \hat{\beta}_n\}$.

If we denote by β_0 the true value of β from equation (2.1) then we note that

$$\hat{E}_n\{D^*(g_{opt}), \beta_0\} = n^{-1} \sum_{i=1}^{n} h(X_i, \beta_0),$$

where

$$h(X_i, \beta) = \mu_i(1, \beta) \mathrm{I}\{\mu_i(1, \beta) \leq \mu_i(0, \beta)\} + \mu_i(0, \beta) \mathrm{I}\{\mu_i(1, \beta) > \mu_i(0, \beta)\}, \tag{2.3}$$

is an empirical average of i.i.d random variables which converges to and has mean

$$E\{h(X_i, \beta_0)\} = E\{D^*(g_{opt})\}.$$

Because of the indicator functions, $h(X_i, \beta)$ may not be differentiable for all β, hence we can not be guaranteed of the usual Taylor series expansion to derive the influence function for $\hat{E}_n\{D^*(g_{opt}), \hat{\beta}_n\}$. Instead we note that

$$\begin{aligned} n^{1/2} \left[\hat{E}_n\{D^*(g_{opt}), \hat{\beta}_n\} - E\{D^*(g_{opt})\}\right] = & \\ n^{1/2} \left[\hat{E}_n\{D^*(g_{opt}), \beta_0\} - E\{D^*(g_{opt})\}\right] + & \qquad (2.4) \\ n^{1/2} \left[\hat{E}_n\{D^*(g_{opt}), \hat{\beta}_n\} - \hat{E}_n\{D^*(g_{opt}), \beta_0\}\right]. & \qquad (2.5) \end{aligned}$$

Because of (2.3), equation (2.4) is equal to

$$n^{-1/2} \sum_{i=1}^{n} [h(X_i, \beta_0) - E\{h(X_i, \beta_0)\}] \tag{2.6}$$

where $E\{h(X_i, \beta)\} = E\{D^*(g_{opt}), \beta\}$ and $E\{h(X_i, \beta_0)\} = E\{D^*(g_{opt})\}$. Under suitable

regularity conditions equation (2.5) can be written as

$$n^{1/2}\left[\hat{E}_n\{D^*(g_{opt}),\hat{\beta}_n\}-\hat{E}_n\{D^*(g_{opt}),\beta_0\}\right]= \frac{\partial E\{D^*(g_{opt}),\beta_0\}}{\partial\beta^T}n^{1/2}(\hat{\beta}_n-\beta_0)+o_p(1).$$

Remark - We wish to note that whenever $\mu(1,x,\beta_0)=\mu(0,x,\beta_0)$ then, by definition, $h(x,\beta)\geq h(x,\beta_0)$ for all β in a neighborhood of β_0, which for most models would imply that $h(x,\beta)$ is not differentiable as a function in β at $\beta=\beta_0$. Consequently, if $P\{\mu(1,X,\beta_0)=\mu(0,X,\beta_0)\}>0$, then $E\{D^*(g_{opt}),\beta\}$ would not be differentiable in β at $\beta=\beta_0$ and hence the asymptotic theory would no longer hold. Therefore, we must be careful to avoid such situations. For example, if the strong null hypothesis were true; i.e., suppose $\mu(1,x,\beta_0)=\mu(0,x,\beta_0)$ for all values of x, then it doesn't matter which treatment is given to any patient and the issue of finding the optimal treatment regime is no longer of interest. Therefore, to avoid this difficulty, we suggest that one first consider testing the null hypothesis for no treatment difference and only if there is strong evidence of treatment difference should one continue the exercise of deriving the optimal treatment regime and its attributable benefit. Recall that standard results for the maximum likelihood estimator can be used to show that $n^{1/2}(\hat{\beta}_n-\beta_0)=n^{-1/2}\sum_{i=1}^{n}\psi_{\hat{\beta}_n}(O_i,\beta_0)+o_p(1)$, where $\psi_{\hat{\beta}_n}(O_i)$ was defined in (1.11).

Combining these results, we have

$$n^{1/2}\left[\hat{E}_n\{D^*(g_{opt}),\hat{\beta}_n\}-E\{D^*(g_{opt})\}\right]=$$

$$n^{-1/2}\sum_{i=1}^{n}\left(h(X_i,\beta_0)-E\{h(X_i,\beta_0)\}+[\partial E\{D^*(g_{opt}),\beta_0\}/\partial\beta^T]\psi_{\hat{\beta}_n}(O_i,\beta_0)\right)+o_p(1).$$

Let us denote the i^{th} term in this summation; i.e., the influence function of $\hat{E}_n\{D^*(g_{opt}),\hat{\beta}_n\}$, as $q_i=q(O_i,\beta_0)$. Therefore, the estimated influence function is given by

$$\hat{q}_i=\left\{h(X_i,\hat{\beta}_n)-n^{-1}\sum_{j=1}^{n}h(X_j,\hat{\beta}_n)\right\}+\left[\partial\hat{E}\{D^*(g_{opt}),\beta_0\}/\partial\beta^T\right]\psi_{\hat{\beta}_n}(O_i,\hat{\beta}_n), \qquad (2.7)$$

where the gradient $\partial E\{D^*(g_{opt}),\beta_0\}/\partial\beta^T$ is estimated using numerical derivatives; namely, where the j^{th} element of the vector $\partial\hat{E}\{D^*(g_{opt}),\beta_0\}/\partial\beta^T$ is obtained by a numerical derivative,

$$\partial\hat{E}\{D^*(g_{opt}),\beta_0\}/\partial\beta^T=[\hat{E}_n\{D^*(g_{opt}),\hat{\beta}_n+\varepsilon_j 1_j\}-\hat{E}_n\{D^*(g_{opt}),\hat{\beta}_n-\varepsilon_j 1_j\}]/2\varepsilon_j,$$

$j = 1, ..., k$, where k is the dimension of the parameter β, and 1_j is a k-dimensional vector where the j^{th} element is 1 and all the other elements are 0. We chose $\varepsilon_j = \hat{\sigma}(\hat{\beta}_j)/100$, where $\hat{\sigma}(\hat{\beta}_j)$ is the estimated standard error of the j^{th} element of β; however, we found that the result was insensitive to the choice of ε_j as long as ε_j was small.

Now that we have derived the influence function for $\hat{E}_n\{D^*(g_{opt}), \hat{\beta}_n\}$, we note that the i^{th} influence function for $\log[\hat{E}_n\{D^*(g_{opt}), \hat{\beta}_n\}] - \log(\bar{D}_n)$, call it ψ_i, can be written as

$$\psi_i = \frac{h(X_i, \beta_0) - E\{h(X_i, \beta_0)\} + [\partial E\{D^*(g_{opt}), \beta_0\}/\partial \beta^T]\psi_{\hat{\beta}_n}(O_i, \beta_0)}{E\{D^*(g_{opt}), \beta_0\}} - \frac{D_i - E(D)}{E(D)}.$$

If we further posit a regression model such as logistic regression,

$$\text{logit}[P(D_i = 1|X_i, E_i)] = \beta^T f(X_i, E_i) = \beta^T f_i$$

where f_i is a $k \times 1$ vector, then note that

$$\mu_i(\beta) = P(D_i = 1|X_i, E_i) = 1 - \{1 + \exp(\beta^T f_i)\}^{-1}.$$

Substituting, we have

$$\psi_{\hat{\beta}_n}(O_i, \beta) = \left(E\left[\frac{f_i f_i^T \exp(\beta^T f_i)}{\{1 + \exp(\beta^T f_i)\}^2} \right] \right)^{-1} f_i \left\{ D_i - \frac{\exp(\beta^T f_i)}{1 + \exp(\beta^T f_i)} \right\}.$$

Likewise we should point out that if we denote $f_i(0) = f(X_i, 0)$ and $f_i(1) = f(X_i, 1)$ then

$$h(X_i, \beta) = \min\left(\frac{\exp\{\beta^T f_i(0)\}}{[1 + \exp\{\beta^T f_i(0)\}]}, \frac{\exp\{\beta^T f_i(1)\}}{[1 + \exp\{\beta^T f_i(1)\}]} \right).$$

For this scenario we will estimate the influence function for the i^{th} individual, ψ_i, from observed data with the quantity

$$\hat{\psi}_i = \frac{\Delta_{1i} + \Delta_{2i}}{\Delta_3} - \Delta_{4i},$$

where

$$\begin{aligned}
\Delta_{1i} &= h(X_i, \hat{\beta}_n) - n^{-1}\sum_{i=1}^{n} h(X_i, \hat{\beta}_n), \\
\Delta_{2i} &= \frac{\partial \hat{E}\{D^*(g_{opt}), \hat{\beta}_n\}}{\partial \hat{\beta}_n^T}\left(n^{-1}\sum_{i=1}^{n}\left[\frac{f_i f_i^T \exp(\hat{\beta}_n^T f_i)}{\{1+\exp(\hat{\beta}_n^T f_i)\}^2}\right]^{-1}\right) f_i\left\{D_i - \frac{\exp(\hat{\beta}_n^T f_i)}{1+\exp(\hat{\beta}_n^T f_i)}\right\}, \\
\Delta_3 &= n^{-1}\sum_{i=1}^{n} h(X_i, \beta), \\
\Delta_{4i} &= \frac{D_i - n^{-1}\sum_{i=1}^{n} D_i}{n^{-1}\sum_{i=1}^{n} D_i}.
\end{aligned}$$

2.2.1 Confidence intervals

Now that we have derived an estimated influence function for $\log[\hat{E}_n\{D^*(g_{opt}), \hat{\beta}_n\}] - \log(\bar{D}_n)$, the estimated variance, using the sandwich variance estimator, is given by

$$\begin{aligned}
\hat{V} &= \hat{Var}(\log[\hat{E}_n\{D^*(g_{opt}), \hat{\beta}_n\}] - \log(\bar{D}_n)) \\
&= n^{-1}\sum_{i=1}^{n}\left(\frac{\Delta_{1i}+\Delta_{2i}}{\Delta_3} - \Delta_{4i}\right)\left(\frac{\Delta_{1i}+\Delta_{2i}}{\Delta_3} - \Delta_{4i}\right)^T.
\end{aligned}$$

We consider two different methods for constructing $(1-\alpha)$ confidence intervals for AB_{opt}. The first involves exponentiating the $(1-\alpha)^{th} * 100\%$ confidence interval for $\log[E\{D^*(g_{opt}), \beta_0\}] - \log\{E(D)\}$ to obtain the following interval

$$1 - \exp\left[\log\left(\sum_{i=1}^{n} h(X_i, \hat{\beta}_n) / \sum_{i=1}^{n} D_i\right) \pm z_{\alpha/2}\sqrt{\hat{V}}\right],$$

where $z_{\alpha/2}$ denotes the $(1-\alpha/2)^{th}$ quantile of a standard normal distribution.
While the second confidence interval is a by-product of the delta theorem, for which we have

$$\hat{AB}_{opt} \pm z_{\alpha/2}\left(\sum_{i=1}^{n} h(X_i, \hat{\beta}_n) / \sum_{i=1}^{n} D_i\right)\sqrt{\hat{V}}.$$

It is not immediately clear if one interval is superior to the other, though it has been suggested that the back-transformed intervals are superior. We will explore this through simulation in the next section.

2.3 Simulation Studies

We report results of several simulations, each based on 10000 Monte Carlo data sets. For simplicity we consider data in the simplest case, of the form $O_i = (D_i, E_i, X_i)$ where D_i is binary disease status, E_i is treatment, and X_i will be one potentially confounding covariate generated from a $N(0,1)$ distribution. For each given X_i we generated a Bernoulli treatment indicator E_i from the logistic regression model:

$$\text{logit}\{P(E_i = 1|X_i)\} = \alpha_0 + \alpha_1 X_i, \tag{2.8}$$

and a Bernoulli disease indicator D_i from the logistic regression model:

$$\text{logit}\{P(D_i = 1|E_i, X_i)\} = \beta_0 + \beta_1 E_i + \beta_2 X_i + \beta_3 E_i X_i. \tag{2.9}$$

Thus different choices of $\beta = \{\alpha_0, \alpha_1, \beta_0, \beta_1, \beta_2, \beta_3\}$ will lead to different $\omega = \{P(E = 1), P(D = 1), P(D^*(0) = 1), P(D^*(1) = 1), P(D(g_{opt}) = 1)\}$, therefore different combinations of the model parameters allow us to study the different items of interest. From model (2.9), one can see that the parameter β_3, the parameter associated with the interaction, will be essential for finding scenarios where there is no one dominant treatment.

Once the data were generated, estimates of $(\beta_0, \beta_1, \beta_2, \beta_3)$ were found for each Monte Carlo dataset using SAS logistic procedure (estimates and standard errors were output using the ODS delivery system). The generated data and logit model estimates were then input into SAS IML where $\hat{AB}_{opt}, \Delta_{1i}, \Delta_{2i}, \Delta_3, \Delta_{4i},\ i = 1, \ldots, n$, and confidence intervals were calculated for each Monte Carlo dataset.

One important point in the influence function calculation is in estimating the derivative $\partial E\{D^*(g_{opt}), \beta\}/\partial \beta^T$ with a numeric approximation. Recall that we chose ε_j to be a 100^{th} of a standard error of each parameter estimate. Under a wide variety of examples (including those listed in table 2.1), $h(X, \hat{\beta}_n)$ was found to be monotone in a neighborhood of size $.01\hat{\sigma}(\hat{\beta}_n)$ for each of the parameters, thus making the numeric approximation possible. In the simulation 1 scenario where there is no interaction ($\beta_3 = 0$), treatment 0 dominates treatment 1 and AB_{dom} (the attributable benefit of the dominant treatment) and AB_{opt} are the same. When one treatment dominates, a closed form derivative can be calculated and simulations show that for simulation 1, the numeric approximation and the closed form derivative agreed up to 4 decimal places.

Table 2.1 illustrates 8 different examples with different combinations of model parameters. We depict the true β and ω values along with estimates of $\hat{AB}_{opt}$, Monte Carlo standard error, delta-theorem standard error and confidence intervals for both back-transformed and delta-theorem methods averaged over all 10000 Monte Carlo samples. Sample sizes of $n = 1000$ were used in all examples; however, we found that the coverage probabilities were accurate in smaller sample sizes, except in those cases where the probabilities in ω fell below 10%.

Examining table 2.1 we see that the Monte Carlo average of the estimates for $\hat{AB}_{opt}$ were indeed very close to the true values as was the Monte Carlo standard error and the average of the delta-theorem standard errors. Both sets of confidence intervals provided very good coverage with probabilities around 0.95. In most cases the back-transformed interval had better coverage than its delta-theorem counterpart.

From these examples we see the importance of AB_{opt} as a measure of a regime's benefit when there are even small interactions between treatment and confounders. We also want to point out that in simulations 4 and 8 the observed proportion of deaths were smaller than either regimes g_0 and g_1 which resulted in a negative AB_{dom}. In these particular simulation scenarios the better treatment was more likely to be given as a function of X and hence the observed proportion of deaths was smaller than a strategy which gave everyone either treatment alone. This further illustrates that AB_{opt} is a more appropriate measure of public health benefit associated with treatment strategies.

In addition to the simulations listed in table 2.1 additional simulations were performed to look at various aspects of model and regime misspecification. Simulations with multiple covariates had similar results to the single covariate case in showing the effectiveness of AB_{opt} as a measure of public health interest. Model overfitting was studied in the multiple covariate case and confidence intervals from overfit models maintained good coverage for the true AB_{opt}, as expected the overfit models did lose efficiency versus the correct model.

One particular simulation of note is when simulation 2 was ran but the model was underfitted by not having an important interaction term. That is, the data were generated as specified in table 2.1 simulation 2, but the model that was fit was $\text{logit}\{P(D_i = 1|E_i, X_i)\} = \beta_0 + \beta_1 E_i + \beta_2 X_i$. In this particular scenario, the proportion of deaths from the dominant treatment was only slightly larger than the optimal treatment regime resulting in similar AB's (i.e., $AB_{dom} = .577$, $AB_{opt} = .588$). However, when the incorrect model was fit

Table 2.1: Simulation Results; BT represent back-transformed confidence intervals while DT represents delta-theorem based confidence intervals

Simulation	1	2	3	4	5	6	7	8
n	1000	1000	1000	1000	1000	1000	1000	1000
Reps	10000	10000	10000	10000	10000	10000	10000	10000
α_0	-1.25	-1.25	-1.25	-1.25	-1.25	-1.25	-1	1
α_1	2	2	2	2	0	-2	-1	3
β_0	-2.5	-2.5	-2.5	-2.5	-2.5	-2.5	1	-3
β_1	1.25	1.25	1.25	1.25	1.25	1.25	1	-1
β_2	1.1	1.1	1.1	1.1	1.1	1.1	1	1
β_3	0	2	-1	-2	-2	-2	-1	-3
$P(E=1)$	0.318	0.318	0.318	0.318	0.223	0.318	0.303	0.613
$P(D=1)$	0.184	0.263	0.122	0.09	0.143	0.222	0.786	0.019
$P(D^*(0)=1)$	0.111	0.111	0.111	0.111	0.111	0.111	0.697	0.069
$P(D^*(1)=1)$	0.267	0.363	0.223	0.2545	0.2545	0.2545	0.881	0.068
$P(D(g_{opt})=1)$	0.111	0.108	0.1	0.068	0.068	0.068	0.691	0.014
AB_{dom}	0.396	0.577	0.094	-0.231	0.223	0.5	0.113	-2.475
AB_{opt}	0.396	0.588	0.184	0.244	0.523	0.693	0.121	0.298
$\hat{AB}_{opt}$	0.3953	0.5897	0.1982	0.2548	0.5277	0.6934	0.1251	0.3407
$\hat{SE}(\hat{AB}_{opt})_{mc}$	0.0799	0.0584	0.0605	0.047	0.0538	0.0442	0.0168	0.1049
$\hat{SE}(\hat{AB}_{opt})_{dt}$	0.0786	0.0578	0.0593	0.048	0.0536	0.0445	0.0169	0.1061
BT Lower 95%	0.2195	0.4587	0.072	0.1549	0.4098	0.5924	0.0914	0.0896
BT Upper 95%	0.5310	0.6885	0.3055	0.3421	0.6218	0.7692	0.1574	0.5161
BT Coverage	0.9415	0.9428	0.9477	0.9529	0.9492	0.9473	0.9579	0.9557
DT Lower 95%	0.2413	0.4764	0.082	0.1612	0.4226	0.6062	0.092	0.1328
DT Upper 95%	0.5493	0.703	0.3144	0.3478	0.6328	0.7806	0.1581	0.5486
DT Coverage	0.9384	0.9352	0.942	0.9479	0.9418	0.944	0.9568	0.9290

without the interaction term, then such a model can only choose a single treatment to be optimal and, in this case, the estimated attributable benefit for that dominant treatment had a mean of .458 rather than the true .577. Thus a misspecified model not only cannot find the optimal treatment regime, but it also results in a biased estimator for the treatment regime it does choose as optimal.

2.4 Example

Our motivating example comes from the Duke Databank for Cardiovascular Diseases (DDCD), a large database housed at the Duke University Medical Center (DUMC) with nearly 40 years of baseline and follow-up data on patients undergoing cardiac catheterization. In this manuscript, we are focused on the subset of patients with a previous CABG surgery and requiring a later catheterization due to continued symptoms. The only treatment options for the subgroup of patients studied are either medical therapy (MED) or some combination of medical therapy and percutaneous coronary intervention (PCI), also known as angioplasty. A question of clinical interest is whether PCI provides some benefit (e.g., lower 1-year mortality rates) above optimal medical therapy for either the overall study population or some subset defined by clinical characteristics.

In studying this question we will explore the effects that covariates play as potential confounders. To see the benefit of using our estimator we will compare estimators from models with zero, one, and several confounding variables. The goal of which is to explore the impact of an optimal treatment strategy while simultaneously examining the effects of confounding. The benefit of the optimal treatment regime will be estimated by the proportion of one-year deaths that could have been prevented if that optimal regime had been followed. The study population includes 3856 individuals with a follow-up catheterization between the years 1986-2001 and a history of CABG surgery.

Table 2.2 provides the number of deaths within 1 year of the follow-up catheterization and a comparison of patient characteristics between the PCI and MED treatment groups. We see a larger proportion of first year deaths for those with medical therapy alone (7.83%) than a combination of medical therapy and angioplasty (5.27%). We will define treatment in terms of a patient obtaining the poorer naive treatment. That is $E = \{0$ if the patient received angioplasty and medical therapy, 1 if the patient received only medical therapy$\}$, we do this so that the optimal treatment regime defaults to the MED group. In

Table 2.2: Treatment versus Death within One Year

Treatment	No Death	Death	Total
PCI + Medical Therapy	1295	72	1367
Medical Therapy	2294	195	2489
Total	3589	267	3856

trying to find the optimal treatment regime we used logistic regression models to predict the probability of death at or before 1 year. In this example, we will assume that the optimal treatment regime will come from the "best" model of one year death rates based on treatment and covariates.

For this discussion we will consider many different models, starting with the simplest model: first year deaths modeled as a function of treatment with no confounding variables. Table 2.3 has several different logistic regression outputs, starting with treatment as the sole predictor in model 1. With the complete data we have an estimate of the probability of first year death $\hat{P}(D = 1) = 0.0692$. Under this model the treatment regime would be g_0, because without interactions the model will favor a combination of angioplasty and medical therapy. From the data we calculate an estimate of $\hat{AB}_{opt} = 0.2393$ with confidence intervals $\{0.0787, 0.3720\}$ (back-transformed method) and $\{0.0936, 0.3851\}$ (delta method).

To begin to see the effect of confounding, model 2 introduces a statistically significant covariate ($p < 0.05$). This model again favors g_0 as the regime of choice and we find $\hat{AB}_{opt} = 0.1572$ with confidence intervals $\{-0.0197, 0.3034\}$ (back-transformed method) and $\{-0.0034, 0.3177\}$ (delta method). We see an 8% drop in risk reduction with the addition of a single confounder, and note the agreement between the AB_{opt} confidence intervals and logistic model p-value for treatment effect. For model 3, forward and backward selection regression techniques were used to arrive at the "best" model for this data. Ejection fraction values were missing for 561 patients, 420 of which had angioplasty and medical therapy as their treatment. In order to keep all observations in the analysis an indicator variable was defined for presence or absence of ejection fraction value, then an interaction term was used in the model to determine the effect of an ejection fraction measurement. Note that if a patient did not have an ejection fraction measurement then the indicators value is zero. This is an important aspect to the results because this model now has different confounders working in both directions; that is some confounders result in a higher probability of dying

within the first year of treatment, while others lower that probability.

Model 3 has an estimate $\hat{AB}_{opt}$ equal to 0.2157 with 95% confidence intervals $\{0.0356, 0.3621\}$ (back-transformed method) and $\{0.0536, 0.3777\}$ (delta method). Therefore the "best" model for this data has AB_{opt} estimates very close to the naive model, but there would have been no way to know apriori that the confounders would have worked in opposing directions. Model 3 again favors the optimal treatment regime where every patient is treated with a combination of angioplasty and medical therapy. To examine the effect of extraneous variables in the model we consider model 4 which has a non-significant interaction term added (interacting treatment with congestive heart failure status). Model 4 had an estimate of $\hat{AB}_{opt} = 0.1950$ with confidence intervals $\{.0066, 0.3476\}$ (back-transformed method) and $\{0.0257, 0.3642\}$ (delta method). So we see that adding an insignificant variable to the model lowered $\hat{AB}_{opt}$ by roughly 2% while increasing the delta theorem standard error. Therefore, we conclude from this analysis that all patients should be treated with a combination of angioplasty and medical therapy.

Table 2.3: Logistic Regression Models, analysis performed with SAS 9.1 logistic procedure.

Response: Death Within First Year

Model 1	DF	Estimate	SE	Wald χ^2	P-value
Intercept	1	-2.8896	0.1211	569.52	$< .0001$
Treatment	1	0.4245	0.1422	8.91	0.0028
Model 2	**DF**	**Estimate**	**SE**	**Wald χ^2**	**P-value**
Intercept	1	-3.2386	0.1345	579.68	$< .0001$
Treatment	1	0.2654	0.1448	3.36	0.0668
Congestive Heart Failure	1	1.0518	0.1301	65.36	$< .0001$
Model 3	**DF**	**Estimate**	**SE**	**Wald χ^2**	**P-value**
Intercept	1	-4.8861	0.4385	124.16	$< .0001$
Treatment	1	0.3999	0.1601	6.24	0.0125
Congestive Heart Failure	1	0.6040	0.1420	18.10	$< .0001$
Number of Disease Vessels	1	0.3487	0.1013	11.86	0.0006
Mitral Insufficiency	1	0.2744	0.0619	19.64	$< .0001$
Pulse	1	0.0182	0.0042	18.81	$< .0001$
I(EJFX Present)	1	0.2215	0.2744	0.65	0.4194
Ejection Fraction*I(EJFX Present)	1	-0.0231	0.0049	22.08	$< .0001$
Model 4	**DF**	**Estimate**	**SE**	**Wald χ^2**	**P-value**
Intercept	1	-4.9472	0.4446	123.82	$< .0001$
Treatment	1	0.5454	0.2163	6.36	0.0117
Congestive Heart Failure	1	0.8205	0.2545	10.40	0.0013
Number of Disease Vessels	1	0.3468	0.1014	11.71	0.0006
Mitral Insufficiency	1	0.2765	0.0620	19.90	$< .0001$
Pulse	1	0.0180	0.0042	18.32	$< .0001$
I(EJFX Present)	1	0.2225	0.2738	0.66	0.4165
Ejection Fraction*I(EJFX Present)	1	-0.0235	0.0049	22.68	$< .0001$
Treatment*CHF	1	-0.3048	0.2987	1.04	0.3075

Chapter 3

Doubly Robust Estimator

3.1 Introduction

So far we have limited our discussion to talking about estimating two types of attributable benefit, the first is AB_{opt}, the benefit associated with the optimal treatment regime. The second is AB_{dom}, the benefit associated with giving all patients the dominant treatment. The definition of an optimal treatment regime is based on correctly specifying the outcome regression (disease) model as a function of covariates and treatment. However, as we pointed out in the previous section, even a misspecified regression model may lead to a treatment regime that is beneficial even if it is not optimal. Therefore it would be important to be able to estimate the attributable benefit unbiasedly for such treatment regimes.

In this section we begin by considering the attributable benefit associated with an arbitrary binary treatment regime $g(X)$, which will be defined as

$$AB_g = 1 - \frac{P[D^*\{g(X)\} = 1]}{P(D = 1)}.$$

The focus of this thesis is on estimating $P[D^*\{g(X)\} = 1]$ and so far we have put forth one method that only considers outcome regression models to estimate the unknown quantity from observed data $O_i = (D_i, E_i, X_i)$. For a given treatment regime, our current estimate of $P[D^*\{g(X)\} = 1]$ is $n^{-1}\sum_{i=1}^{n}[\mu_i(1, \hat{\beta}_n)g(X_i) + \mu_i(0, \hat{\beta}_n)\{1 - g(X_i)\}]$. The work done in the last chapter shows that the regression based estimator is consistent for the quantity of interest and asymptotically normal, given the underlying assumption that we have the appropriate model. That is to say that the entire methodology is predicated on having the

correct statistical model. We have already seen through simulation that if we do not have the correct model there will be biased estimates of AB_g. We now turn our attention to deriving an estimator that may offer protection from such misspecification.

3.1.1 Inverse Propensity Weights

One way of deriving an unbiased estimator for $P[D^*\{g(X)\} = 1]$ without having to specify an outcome regression model for $\mu(X_i, E_i) = P(D = 1|E, X)$ is by the use of an inverse propensity score weighted estimator, which we will refer to as an inverse propensity estimator for short. The propensity score is defined as the probability of receiving a treatment given covariates; i.e., $P(E = 1|X)$. We start by positing a statistical model where

$$P(E = 1|X) = \pi_E(X, \alpha). \tag{3.1}$$

Because E is a binary variable for treatment, a natural choice is to use logistic regression models which allow us to study the impact of covariates and their interaction. For model (3.1) the parameter α can be estimated using maximum likelihood; that is, the maximum likelihood estimator $\hat{\alpha}_n$ would be obtained by maximizing (in α) $\prod_{i=1}^{n}\{\pi_E(X, \alpha)\}^{E_i}\{1 - \pi_E(X, \alpha)\}^{1-E_i}$.

Using standard results from causal inference [8], the inverse propensity score weighted estimator for $P\{D^*(1) = 1\}$ is given as

$$\hat{P}_n\{D^*(1) = 1\} = n^{-1}\sum_{i=1}^{n}\frac{E_i D_i}{\pi_E(X, \hat{\alpha}_n)}$$

where $\pi_E(X, \hat{\alpha}_n)$ is the estimate of $\pi_E(X, \alpha)$ at the maximum likelihood estimator $\hat{\alpha}_n$. What we want is an estimate of $P[D^*\{g(X)\} = 1]$, so we will define the following indicator function $C(g)$ that will denote agreement between the treatment observed and the treatment suggested by some strategy g. That is, $C(g) = \{1$ if $g(X) = t$ and patient received treatment $t, 0$ otherwise$\}$ (where $t = 0, 1$). Specifically, we can write $C(g)$ as

$$C(g) = g(X)E + \{1 - g(X)\}(1 - E).$$

In order to estimate $P[D^*\{g(X)\} = 1]$ with an inverse propensity estimator we need $P\{C(g) = 1|X\}$ which is

$$\begin{aligned}\pi_{C(g)}(X_i) &= P\{C(g) = 1|X_i\} \\ &= \{\pi_E(X_i, \alpha)\}g(X_i) + \{1 - \pi_E(X_i, \alpha)\}\{1 - g(X_i)\}. \end{aligned} \tag{3.2}$$

Recall that

$$D^*\{g(X)\} = D^*(1)g(X) + D^*(0)\{1 - g(X)\},$$

and the assumption of no unmeasured confounders states that E is conditionally independent of $\{D^*(0), D^*(1)\}$ given X. As $D^*\{g(X)\}$ is a function of $\{D^*(0), D^*(1), X\}$ and $C\{g(X)\}$ is a function of E, X then no unmeasured confounders implies $C\{g(X)\}$ is conditionally independent of $D^*\{g(X)\}$ given X. This conditional independence will play an important role in the development of new estimators for $P[D^*\{g(X)\} = 1]$.

We therefore propose the inverse propensity estimator of $P[D^*\{g(X)\} = 1]$ to be

$$\hat{P}_n[D^*\{g(X)\} = 1] = n^{-1} \sum_{i=1}^{n} \frac{C\{g(X_i)\}D_i}{\pi_{C(g)}(X_i, \hat{\alpha}_n)}, \tag{3.3}$$

where $\pi_{C(g)}(X_i, \hat{\alpha}_n)$ is the estimate of $\pi_{C(g)}(X, \alpha)$ for the i^{th} patient at the maximum likelihood estimator $\hat{\alpha}_n$. The question now is whether this estimator is consistent for the quantity of interest? Under the assumption that the model for propensity score, $\pi_E(x, \alpha)$ is correctly specified; i.e., $P(E = 1|X) = \pi_0(X) = \pi_E(X, \alpha)$ for some α (which we will denote by α_0), then under SUTVA we have $C\{g(X_i)\}D_i = C\{g(X_i)\}D^*\{g(X_i)\}$ and $C\{g(X)\}$ is conditionally independent of $D^*\{g(X)\}$ given X then (3.3) converges to

$$\begin{aligned}
E\left[\frac{C\{g(X)\}D}{\pi_{C(g)}(X, \alpha_0)}\right] &= E\left[\frac{C\{g(X)\}D^*\{g(X)\}}{\pi_{C(g)}(X, \alpha_0)}\right] \\
&= E\left(E\left[\frac{C\{g(X)\}D^*\{g(X)\}}{\pi_{C(g)}(X, \alpha_0)}|X, D^*\{g(X)\}\right]\right) \\
&= E\left(\frac{D^*\{g(X)\}}{\pi_{C(g)}(X, \alpha_0)} E\left[C\{g(X)\}|X, D^*\{g(X)\}\right]\right) \\
&= E\left(\frac{D^*\{g(X)\}}{\pi_{C(g)}(X, \alpha_0)} E\left[C\{g(X)\}|X\right]\right) \\
&= E\left(\frac{D^*\{g(X)\}}{\pi_{C(g)}(X, \alpha_0)} \pi_{C(g)}(X, \alpha_0)\right)
\end{aligned}$$

If we now assume $0 < P(E = 1|X) < 1$ for all values of X (so $\pi_{C(g)}(X, \alpha_0)/\pi_{C(g)}(X, \alpha_0) = 1$) then

$$E\left[\frac{C\{g(X)\}D}{\pi_{C(g)}(X, \alpha)}\right] = E\left[D^*\{g(X)\}\right] = P[D^*\{g(X)\} = 1].$$

Thus we have a way to estimate $P[D^*\{g(X)\} = 1]$ that is free of any potential misspecification of $P(D = 1|E, X)$, but inverse propensity estimators have some well documented problems of their own; most notably is when, for some value of X, $P(E = 1|X)$ is very close

to 0 or 1. If $P(E = 1|X)$ is very small then the corresponding inverse weight is very large and often leads to unstable estimators. From a real-world standpoint, if this conditional probability reaches 0 or 1 then we cannot assess the relative contribution of treatments 1 or 0 for individuals with covariate values where one or the other treatments were given with certainty.

Another issue in using such an estimator in this format is that we may have indeed dealt with misspecification of $P(D = 1|E, X)$, but we now have the same potential problem in misspecifying $P(E = 1|X)$. Lastly, if the outcome regression model were correctly specified, then the outcome regression estimator proposed in the previous chapter is a maximum likelihood estimator for $P[D^*\{g(X)\} = 1]$ and hence should perform better than an estimator that ignores their dependence.

3.2 A Doubly Robust Estimator for $P[D^*\{g(X)\} = 1]$

It is well known that one can improve the performance of inverse propensity estimators by augmenting them with additional terms that involve the outcome (disease) regression model resulting in the so-called doubly robust estimator [11]. Generally speaking, a doubly robust estimator for $P[D^*\{g(X)\} = 1]$ will augment the inverse propensity estimator with a term involving $P(D = 1|E, X)$ in such a way that the new estimator is consistent for $P[D^*\{g(X)\} = 1]$ as long as models for either $P(E = 1|X)$ or $P(D = 1|E, X)$ are correct. Named doubly robust for its dual layers of protection from misspecification, such an estimator would reduce bias from potentially misspecified models and uses information gained from the relationship between E and X as well as between D and (E, X) [9] [11].

In laying the groundwork for discussing model misspecification note that there exist some true values for $P(E = 1|X)$ and $P(D = 1|E, X)$ which we will refer to as $\pi_0(X)$ and $\mu_0(E, X)$ respectively. In developing the methodology, statistical models are posited for the propensity score $\pi_E(X, \alpha)$ and for the outcome regression $\mu(E, X, \beta)$ in terms of unknown parameters α and β, respectively, which may or may not be correctly specified. We will estimate these parameters from data using maximum likelihood estimators which we denote by $\hat{\alpha}_n$ and $\hat{\beta}_n$, respectively. Note that under suitable regularity conditions, the resulting estimators will converge to some constants; i.e., $\hat{\alpha}_n \to \alpha^*$ and $\hat{\beta}_n \to \beta^*$, where α^* and β^* are some constants whether the models are correctly specified or not. If

the estimated propensity model is correctly specified then $\pi_E(X, \alpha^*) = \pi_0(X)$ and if the estimated outcome regression model is correctly specified then $\mu(E, X, \beta^*) = \mu_0(E, X)$. If the propensity score or outcome regression models are misspecified then $\pi_E(X, \hat{\alpha}_n)$ or $\mu(E, X, \hat{\beta}_n)$ will not converge to the desired quantities.

From the posited models we can note that $\pi_{C(g)}(X, \alpha) = P\{C(g) = 1|X\}$ and from chapter (2)

$$P\{D = 1|C(g) = 1, X\} = h(X, \beta) = g(X)\mu(1, X, \beta) + \{1 - g(X)\}\mu(0, X, \beta),$$

where $h(X, \beta)$ is now defined in terms of a general treatment regime g. We can now define

$$\begin{aligned}
\lambda = \lambda(\alpha, \beta) = & \\
& E\left[\frac{C(g)D}{\pi_{C(g)}(X, \alpha)} - \frac{C(g) - \pi_{C(g)}(X, \alpha)}{\pi_{C(g)}(X, \alpha)} h(X, \beta)\right] = \qquad (3.4) \\
& E\left[\frac{C(g)D}{\pi_{C(g)}(X, \alpha)} - \frac{C(g) - \pi_{C(g)}(X, \alpha)}{\pi_{C(g)}(X, \alpha)} \{g(X)\mu(1, \beta) + \{1 - g(X)\}\mu(0, \beta)\}\right]. \qquad (3.5)
\end{aligned}$$

From this we now propose a doubly robust estimator for $P[D^*\{g(X)\} = 1]$ from observed data $O_i = (D_i, E_i, X_i)$ as

$$\begin{aligned}
\hat{\lambda}_n = \quad n^{-1} \sum_{i=1}^{n} \frac{C_i(g)D_i}{\pi_{C_i(g)}(X_i, \hat{\alpha}_n)} \quad & - \frac{C_i(g) - \pi_{C(g)}(X_i, \hat{\alpha}_n)}{\pi_{C(g)}(X_i, \hat{\alpha}_n)} \\
& \{g(X_i)\mu_i(1, \hat{\beta}_n) + \{1 - g(X_i)\}\mu_i(0, \hat{\beta}_n)\}, \qquad (3.6)
\end{aligned}$$

where $\hat{\alpha}_n$ and $\hat{\beta}_n$ are maximum likelihood estimates of α and β respectively. We will now demonstrate that this estimator is indeed double robust.

Claim: $\hat{\lambda}_n$ is doubly robust for $P[D^*\{g(X)\} = 1]$, that is $\hat{\lambda}_n$ converges to $P[D^*\{g(X)\} = 1]$ if either $\pi_E(X, \alpha^*) = \pi_0(X)$ or $\mu(E, X, \beta^*) = \mu_0(E, X)$.

Because, under suitable regularity conditions, $\hat{\alpha}_n \to \alpha^*$ and $\hat{\beta}_n \to \beta^*$ then $\hat{\lambda}_n \to \lambda(\alpha^*, \beta^*)$ [3]. Therefore to prove the claim we must show that $\lambda(\alpha^*, \beta^*)$ is equal to $P[D^*\{g(X)\} = 1]$ if either $\pi_E(X, \alpha^*) = \pi_0(X)$ or $\mu(E, X, \beta^*) = \mu_0(E, X)$. We begin

by noting that under SUTVA

$$\begin{aligned}
\lambda(\alpha^*,\beta^*) &= E\left[\frac{C(g)D}{\pi_{C(g)}(X,\alpha^*)} - \frac{C(g)-\pi_{C(g)}(X,\alpha^*)}{\pi_{C(g)}(X,\alpha^*)}h(X,\beta^*)\right] \\
&= E\left[\frac{C(g)D^*(g)}{\pi_{C(g)}(X,\alpha^*)} - \frac{C(g)-\pi_{C(g)}(X,\alpha^*)}{\pi_{C(g)}(X,\alpha^*)}h(X,\beta^*)\right] \\
&= E\left[\frac{C(g)D^*(g)}{\pi_{C(g)}(X,\alpha^*)} - \frac{C(g)-\pi_{C(g)}(X,\alpha^*)}{\pi_{C(g)}(X,\alpha^*)}h(X,\beta^*) + D^*(g) - D^*(g)\right] \\
&= E\left[D^*(g) + \frac{C(g)-\pi_{C(g)}(X,\alpha^*)}{\pi_{C(g)}(X,\alpha^*)}\{D^*(g)-h(X,\beta^*)\}\right] \\
&= E\{D^*(g)\} + E\left[\frac{C(g)-\pi_{C(g)}(X,\alpha^*)}{\pi_{C(g)}(X,\alpha^*)}\{D^*(g)-h(X,\beta^*)\}\right].
\end{aligned}$$

To prove the claim it is now enough to show that

$$E\left[\frac{C(g)-\pi_{C(g)}(X,\alpha^*)}{\pi_{C(g)}(X,\alpha^*)}\{D^*(g)-h(X,\beta^*)\}\right] = 0$$

if either $\pi_E(X,\alpha^*) = \pi_0(X)$ or $\mu(E,X,\beta^*) = \mu_0(E,X)$. First assume that only $\pi_E(X,\alpha^*) = \pi_0(X)$, then because of the no unmeasured confounders assumption $E\{C(g)|D^*(g),X\} = E\{C(g)|X\} = \pi_{C(g)}(X,\alpha^*)$ and we have

$$\begin{aligned}
&E\left[\frac{C(g)-\pi_{C(g)}(X,\alpha^*)}{\pi_{C(g)}(X,\alpha^*)}\{D^*(g)-h(X,\beta^*)\}\right] = \\
&\quad E\left(E\left[\frac{C(g)-\pi_{C(g)}(X,\alpha^*)}{\pi_{C(g)}(X,\alpha^*)}\{D^*(g)-h(X,\beta^*)\}|D^*(g),X\right]\right) = \\
&\quad E\left(\frac{D^*(g)-h(X,\beta^*)}{\pi_{C(g)}(X,\alpha^*)}\left[E\{C(g)|D^*(g),X\} - \pi_{C(g)}(X,\alpha^*)\right]\right) = 0
\end{aligned}$$

as $E\{C(g)|D^*(g),X\} = \pi_{C(g)}(X,\alpha^*)$.

Now assume that $\mu(E,X,\beta^*) = \mu_0(E,X)$, then again by the no unmeasured confounders assumption $E\{D^*(g)|X,C(g)\} = E\{D^*(g)|X\} = h(X,\beta^*)$ and we have

$$\begin{aligned}
&E\left[\frac{C(g)-\pi_{C(g)}(X,\alpha^*)}{\pi_{C(g)}(X,\alpha^*)}\{D^*(g)-h(X,\beta^*)\}\right] = \\
&\quad E\left(E\left[\frac{C(g)-\pi_{C(g)}(X,\alpha^*)}{\pi_{C(g)}(X,\alpha^*)}\{D^*(g)-h(X,\beta^*)\}|X,C(g)\right]\right) = \\
&\quad E\left(\frac{C(g)-\pi_{C(g)}(X,\alpha^*)}{\pi_{C(g)}(X,\alpha^*)}E\left[D^*(g)-h(X,\beta^*)|X,C(g)\right]\right) = 0
\end{aligned}$$

as $E\{D^*(g)|X, C(g)\} = h(X, \beta^*)$. Thus we have now shown is that as long as we specify a correct model for either propensity score, $P(E = 1|X)$, or outcome regression ,$P(D = 1|E, X)$, then (3.6) is consistent for the quantity of interest.

In estimating α and β from data recall in the previous section we posited a regression model such as logistic regression,

$$\text{logit}[P(D_i = 1|X_i, E_i)] = \beta^T f(X_i, E_i) = \beta^T f_i,$$

where f_i is a $k \times 1$ vector. Therefore, we can write

$$h(X_i, \beta) = g(X_i)\frac{\exp\{\beta^T f_i(0)\}}{[1 + \exp\{\beta^T f_i(0)\}]} + \{1 - g(X_i)\}\frac{\exp\{\beta^T f_i(1)\}}{[1 + \exp\{\beta^T f_i(1)\}]}$$

and consequently the regression based estimator of AB_g; namely,

$$\hat{AB}_g^R = 1 - \left\{\sum_{i=1}^{n} h(X_i, \hat{\beta}_n) / \sum_{i=1}^{n} D_i\right\} \tag{3.7}$$

is a consistent estimator of AB_g^R, where

$$AB_g^R = 1 - [E\{h(X, \beta^*)\}/P(D = 1)].$$

Suppose we also posit a logistic regression model for $P(E = 1|X)$, that is

$$\text{logit}[P(E_i = 1|X_i)] = \alpha^T f^p(X_i) = \alpha^T f_i^p$$

where f_i^p is a $k_p \times 1$ vector. Then the doubly robust estimator of AB_g is given by

$$\hat{AB}_g^{DR} = 1 - \left\{\hat{\lambda}_n / \bar{D}_n\right\} \tag{3.8}$$

which is a consistent estimator of AB_g^{DR}, where

$$AB_g^{DR} = 1 - [E\{\lambda(\alpha^*, \beta^*)\}/P(D = 1)].$$

Note that AB_g^{DR} is equal to AB_g if either the model for propensity score or outcome regression is correctly specified, whereas, AB_g^R is equal to AB_g only if the model for outcome regression is correctly specified.

3.3 Large Sample Properties and Standard Errors

In this section we will demonstrate that $\hat{AB}_g^{DR}$ is asymptotically normal and derive an estimator for its asymptotic variance by finding its influence function. We will impose similar techniques as used in the previous chapter, instead of finding influence functions for α, β, and λ separately; here we will derive a vector of influence functions that we can use to find the large sample variance of the quantity of interest. The ultimate goal will be estimating the influence function for $\log(\hat{\lambda}_n) - \log(\bar{D}_n)$, and constructing confidence intervals similar to those derived in the previous chapter. First we need to find the influence function of $\hat{\lambda}_n$ denoted here by $\psi_{\hat{\lambda}_n}(O_i)$.

Consider the following $(1 + k_p + k) \times 1$ vector

$$m^*(D_i, E_i, X_i, \lambda, \alpha, \beta) = [m_1^*, m_2^*, m_3^*]^T,$$

where m_1^*, m_2^*, m_3^* are defined as

$$\begin{aligned} m_1^* &= \frac{C_i D_i}{\pi_C(X_i, \alpha)} - \frac{C_i - \pi_C(X_i, \alpha)}{\pi_C(X_i, \alpha)} h(X_i, \beta) - \lambda(\alpha, \beta) \\ m_2^* &= f^p(X_i)\{E_i - \pi_E(X_i, \alpha)\} \\ m_3^* &= f(X_i, E_i)\{D_i - \mu(E_i, X_i, \beta)\}. \end{aligned}$$

Note for maximum likelihood estimators $\hat{\alpha}_n$ and $\hat{\beta}_n$ that $n^{-1} \sum_{i=1}^{n} m^*(D_i, E_i, X_i, \hat{\lambda}_n, \hat{\alpha}_n, \hat{\beta}_n) = [0]$ where $[0]$ is a $(1 + k_p + k) \times 1$ vector of zeros, thus m^* is a vector of m-estimators and as such we have already stated that the influence functions for λ, α, β an be expressed as

$$\begin{aligned} \psi(O_i) &= [\psi_{\hat{\lambda}_n}(O_i), \psi_{\hat{\alpha}_n}(O_i), \psi_{\hat{\beta}_n}(O_i)]^T \\ &= \left\{ E\left[\frac{\partial m^*}{\partial \lambda \partial \alpha_0 \ldots \partial \alpha_{k_p} \partial \beta_0 \ldots \partial \beta_k} \right] \right\}^{-1} m^*(D_i, E_i, X_i, \lambda, \alpha, \beta), \end{aligned} \tag{3.9}$$

where $\left[\partial m^* / \partial \lambda \partial \alpha_0 \ldots \partial \alpha_{k_p} \partial \beta_0 \ldots \partial \beta_k\right]$ is the matrix of partial derivatives expressed as

$$\left[\frac{\partial m^*}{\partial \lambda \partial \alpha_0 \ldots \partial \alpha_{k_p} \partial \beta_0 \ldots \partial \beta_k} \right] = \begin{bmatrix} \partial m_1^*/\partial \lambda & \partial m_1^*/\partial \alpha_0 & \ldots & \partial m_1^*/\partial \beta_k \\ \partial m_2^*/\partial \lambda & \partial m_2^*/\partial \alpha_0 & \ldots & \partial m_2^*/\partial \beta_k \\ \partial m_3^*/\partial \lambda & \partial m_3^*/\partial \alpha_0 & \ldots & \partial m_3^*/\partial \beta_k \end{bmatrix}.$$

Recall that m_2^* and m_3^* are column vectors with k_p and k rows accordingly, the partial derivatives of which are simply the partial derivatives of each element in the column vector. When we examine $\left[\partial m^* / \partial \lambda \partial \alpha_0 \ldots \partial \alpha_{k_p} \partial \beta_0 \ldots \partial \beta_k\right]$ we find it has a specific structure, m_2^*

is only a function of α, thus $\partial m_2^*/\partial\lambda = \partial m_2^*/\partial\beta_0 = ... = \partial m_2^*/\partial\beta_k = 0$. Likewise note m_3^* is only a function of β, thus $\partial m_2^*/\partial\lambda = \partial m_2^*/\partial\alpha_0 = ... = \partial m_2^*/\partial\alpha_{k_p} = 0$. Lastly the way we have constructed m_1^* note that $\partial m_1^*/\partial\lambda = -1$. The partial derivatives involving m_1^* are the most complex since m_1^* is a function of λ, α and β. Given the success of numeric derivatives in the last chapter, we will employ a similar technique here and estimate these partial derivatives with numeric derivatives, that is

$$\partial m_1^*(\lambda,\alpha,\beta)/\partial\alpha^T = [m_1^*(\lambda,\hat{\alpha}_n+\tau_r 1_r,\beta) - m_1^*(\lambda,\hat{\alpha}_n-\tau_r 1_r,\beta)]/2\tau_r \quad (3.10)$$

$$\partial m_1^*(\lambda,\alpha,\beta)/\partial\beta^T = [m_1^*(\lambda,\alpha,\hat{\beta}_n+\varepsilon_j 1_j) - m_1^*(\lambda,\alpha,\hat{\beta}_n-\varepsilon_j 1_j)]/2\varepsilon_j, \quad (3.11)$$

where $r = 1, ..., k_p$ and $j = 1, ..., k$ with k_p and k the dimensions of α and β respectively, 1_r is a k_p-dimensional vector where the r^{th} element is 1 and all the other elements are 0. Likewise 1_j is a k-dimensional vector where the j^{th} element is 1 and all the other elements are 0. Similar to the last chapter we chose $\tau_r = \hat{\sigma}(\hat{\alpha}_r)/100$ and $\varepsilon_j = \hat{\sigma}(\hat{\beta}_j)/100$, where $\hat{\sigma}(\hat{\alpha}_r)$ and $\hat{\sigma}(\hat{\beta}_j)$ are the estimated standard errors of the r^{th} and j^{th} elements of α and β respectively.

Now that we have defined $\psi(O_i)$, we are only concerned with the first element so after calculating an estimate for the vector denoted in (3.9) we only need the first element. Recall that the main objective to this is to estimate the influence function for $\log(\hat{\lambda}_n) - \log(\bar{D}_n)$. For the scenario where we have posited logit models stated earlier, we can now estimate this influence function with the quantity

$$\hat{\psi}_i = \frac{\nu_{1i}}{\nu_2} - \nu_{3i},$$

where

$$\begin{aligned}
\nu_{1i} &= \psi_{\hat{\lambda}_n}(D_i, E_i, X_i,) \\
\nu_2 &= n^{-1}\sum_{i=1}^{n} \frac{C_i D_i}{\pi_C(X_i,\hat{\alpha}_n)} - \frac{C_i - \pi_C(X_i,\hat{\alpha}_n)}{\pi_C(X_i,\hat{\alpha}_n)} h(X_i,\hat{\beta}_n) \\
\nu_{3i} &= (D_i - \bar{D}_n)/\bar{D}_n
\end{aligned}$$

with

$$\pi_C(X_i,\hat{\alpha}_n) = \pi_E(X_i,\hat{\alpha}_n)g(X_i) + \{1-\pi_E(X_i,\hat{\alpha}_n)\}\{1-g(X_i)\}.$$

3.3.1 Confidence intervals

Now that we have derived an estimated influence function for $\log(\hat{\lambda}_n) - \log(\bar{D}_n)$, the estimated variance, using the sandwich variance estimator, is given by

$$\hat{V}_{DR} = n^{-1} \sum_{i=1}^{n} \left(\frac{\nu_{1i}}{\nu_2} - \nu_{3i} \right) \left(\frac{\nu_{1i}}{\nu_2} - \nu_{3i} \right)^T .$$

As before we consider two different methods for constructing $(1-\alpha)$ confidence intervals for AB_g^{DR}. The first involves exponentiating the $(1-\alpha)^{th} \times 100\%$ confidence interval for $\log[\hat{\lambda}_n] - \log(\bar{D}_n)$ to obtain the following interval

$$1 - \exp\left[\log\left(\hat{\lambda}_n / \bar{D}_n \right) \pm z_{\alpha/2} \sqrt{\hat{V}_{DR}} \right],$$

where $z_{\alpha/2}$ denotes the $(1-\alpha/2)^{th}$ quantile of a standard normal distribution. While the second confidence interval is a by-product of the delta theorem, for which we have

$$\hat{AB}_{opt}^{DR} \pm z_{\alpha/2} \left(\hat{\lambda}_n / \bar{D}_n \right) \sqrt{\hat{V}_{DR}}.$$

We will explore the merits of each confidence interval through simulation in a later section.

3.3.2 Example

For illustration of the calculation of the influence function for $\hat{\lambda}_n$ let's take a simple example with $O_i = (D_i, E_i, X_i)$ and

$$\text{logit}[P(E=1|X)] = \alpha_0 + \alpha_1 X$$

$$\text{logit}[P(D=1|E,X)] = \beta_0 + \beta_1 E + \beta_2 X$$

Under this scenario we have

$$m^*(D_i, E_i, X_i, \lambda, \alpha, \beta) = \begin{bmatrix} m_1^* \\ \{E_i - \pi_E(X_i, \alpha)\} \\ X_i\{E_i - \pi_E(X_i, \alpha)\} \\ \{D_i - \mu(E_i, X, \beta)\} \\ E_i\{D_i - \mu(E_i, X, \beta)\} \\ X_i\{D_i - \mu(E_i, X, \beta)\} \end{bmatrix}.$$

And

$$\left[\frac{\partial m^*}{\partial\lambda\partial\alpha_0\ldots\partial\alpha_{k_p}\partial\beta_0\ldots\partial\beta_k}\right] = -$$

$$\begin{bmatrix} 1 & \partial m_1^*/\partial\alpha_0 & \partial m_1^*/\partial\alpha_1 & \partial m_1^*/\partial\beta_0 & \partial m_1^*/\partial\beta_1 & \partial m_1^*/\partial\beta_2 \\ 0 & \partial\pi_E(\alpha)/\partial\alpha_0 & \partial\pi_E(\alpha)/\partial\alpha_1 & 0 & 0 & 0 \\ 0 & X_i\partial\pi_E(\alpha)/\partial\alpha_0 & X_i\partial\pi_E(\alpha)/\partial\alpha_1 & 0 & 0 & 0 \\ 0 & 0 & 0 & \partial\mu_i(\beta)/\partial\beta_0 & \partial\mu_i(\beta)/\partial\beta_1 & \partial\mu_i(\beta)/\partial\beta_2 \\ 0 & 0 & 0 & E_i\partial\mu_i(\beta)/\partial\beta_0 & E_i\partial\mu_i(\beta)/\partial\beta_1 & E_i\partial\mu_i(\beta)/\partial\beta_2 \\ 0 & 0 & 0 & X_i\partial\mu_i(\beta)/\partial\beta_0 & X_i\partial\mu_i(\beta)/\partial\beta_1 & X_i\partial\mu_i(\beta)/\partial\beta_2 \end{bmatrix}$$

As stated earlier we can see something of a pattern in $\partial m^*/\partial\lambda\partial\alpha_0\ldots\partial\beta_k$, beyond the first row the partial derivative matrix is made up of blocks of derivatives that involve just α or just β and zero everywhere else. For this example note $\partial\pi_E(\alpha)/\partial\alpha_r$ and $\partial\mu_i(\beta)/\partial\beta_j$ are straightforward logit model calculations (i.e. $\partial\pi_E(\alpha)/\partial\alpha_0 = \pi_E(X_i,\alpha)\{1-\pi_E(X_i,\alpha)\}$), and $\partial m_1^*/\partial\alpha_0, ..., \partial m_1^*/\partial\beta_2$ are estimated in the manner using the numeric derivatives described earlier. Software can be used to estimate this partial derivative for each person in the sample, then average and invert to have an estimate of the expected value of the matrix of partial derivatives. We then take that estimate and multiply by $m^*(D_i, E_i, X_i, \lambda, \alpha, \beta)$ and then pick off the first element to obtain the i^{th} influence function for $\hat{\lambda}_n$. The calculation of the quantities ν_2 and ν_{3i} are straightforward, giving way to an estimate for the influence function of $\log[\hat{\lambda}_n] - \log(\bar{D}_n)$, from which we can construct confidence intervals.

Chapter 4

AB Estimators Under Estimated Treatment Regimes

4.1 Estimating Optimal Treatment Regimes

It was stated in the introduction that this body of work has two goals, to estimate the optimal treatment strategy from observed data and to measure that strategies public health impact by estimating its attributable benefit. So far in our study of a doubly robust estimator of attributable benefit, we have only discussed assessing the benefit of a known treatment strategy. That is to say that if we have a known treatment strategy, we can estimate its attributable benefit with double protection. In clinical practice the treatment strategy that is most important is $g_{opt}(X)$ which in almost every situation is not known and has to be estimated from data. We have already stated that we will estimate the optimal treatment strategy via statistical modeling, recall from earlier that we posited a model

$$P(D = 1|E, X) = \mu(E, X, \beta)$$

and from there we conclude that the optimal treatment strategy would be

$$g_{opt}(X) = \mathrm{I}\{\mu(1, X, \beta) \leq \mu(0, X, \beta)\}.$$

Then we can estimate the optimal treatment strategy by making an estimate of $\mu(E, X, \beta)$, which we have already done from logistic regression models and maximum likelihood.

In chapter (2) we denote the "best" estimate of $\mu(E, X, \beta)$ with $\mu(E, X, \hat{\beta}_n)$ and simulation studies showed that the regression estimator of attributable benefit has good

coverage when the outcome regression model is correctly specified; that is, when the true outcome regression $P(D=1|E,X)=\mu_0(E,X)$ is contained in the model $\mu(E,X,\beta)$. However, as we have demonstrated through simulation studies, the treatment regime that is estimated using a misspecified outcome regression model may often have good attributable benefit even if the outcome regression model is misspecified. The difficulty is that the estimator of attributable benefit for such a regime may be substantially biased if one uses a misspecified outcome regression estimator. This motivates us to use a two-stage method to obtain a treatment regime and an estimator of its attributable benefit. In the first stage we posit an outcome regression model $\mu(E,X,\beta)$ and define the treatment regime of interest as $g(X,\hat{\beta}_n)=\mathrm{I}\{\mu(1,X,\hat{\beta}_n)\leq\mu(0,X,\hat{\beta}_n)\}$, where $\hat{\beta}_n$ is the maximum likelihood estimator of β. In the second stage we estimate the attributable benefit of $g(X,\hat{\beta}_n)$ by using the double robust estimator derived in chapter (3).

4.2 Doubly Robust Estimation of AB from $\mu(E,X,\hat{\beta}_n)$

In chapter (3) we form the doubly robust estimator which will be better equipped for separating the problems of treatment strategy and attributable benefit estimation. So far we know that if we are given a specific treatment strategy (including the optimal treatment regime) then we can estimate its AB with doubly robust protection. Now let's suppose that we don't know the optimal treatment strategy and we estimate it using $\mu(E,X,\hat{\beta}_n)$. Then we write the function $C(E,X,\hat{\beta}_n)$ that denotes agreement between the observed and the treatment suggested by $g(X)$ as

$$C(E,X,\hat{\beta}_n)=\mathrm{I}\{\mu(1,X,\hat{\beta}_n)\leq\mu(0,X,\hat{\beta}_n)\}E+\mathrm{I}\{\mu(1,X,\hat{\beta}_n)>\mu(0,X,\hat{\beta}_n)\}(1-E).$$

Likewise we estimate the function $\pi_{C(g)}(X_i)$ with

$$\begin{aligned}\hat{\pi}_{C(g)}(X_i,\hat{\alpha}_n,\hat{\beta}_n) &= \{\pi_E(X_i,\hat{\alpha}_n)\}\mathrm{I}\{\mu(1,X,\hat{\beta}_n)\leq\mu(0,X,\hat{\beta}_n)\}\\ &+ \{1-\pi_E(X_i,\hat{\alpha}_n)\}\mathrm{I}\{\mu(1,X,\hat{\beta}_n)>\mu(0,X,\hat{\beta}_n)\}. \qquad (4.1)\end{aligned}$$

Recall from chapter (2) that the model based estimate of $h(X_i,\beta)$ (equation (2.3)) is $h(X_i,\hat{\beta}_n)=\mu_i(1,\hat{\beta}_n)\mathrm{I}\{\mu_i(1,\hat{\beta}_n)\leq\mu_i(0,\hat{\beta}_n)\}+\mu_i(0,\hat{\beta}_n)\mathrm{I}\{\mu_i(1,\hat{\beta}_n)>\mu_i(0,\hat{\beta}_n)\}$. From here we estimate $\hat{\lambda}_n$ and AB_{opt} with

$$\hat{\lambda}_n(\hat{\alpha}_n,\hat{\beta}_n)=n^{-1}\sum_{i=1}^{n}\frac{C(E,X,\hat{\beta}_n)D_i}{\hat{\pi}_{C(g)}(X_i,\hat{\alpha}_n,\hat{\beta}_n)}-\frac{C(E,X,\hat{\beta}_n)-\hat{\pi}_{C(g)}(X_i,\hat{\alpha}_n,\hat{\beta}_n)}{\hat{\pi}_{C(g)}(X_i,\hat{\alpha}_n,\hat{\beta}_n)}h(X_i,\hat{\beta}_n) \quad (4.2)$$

$$\hat{AB}_{opt}^{DR} = 1 - \left\{ \sum_{i=1}^{n} \hat{\lambda}_n / \sum_{i=1}^{n} D_i \right\}.$$

Variance estimates and confidence intervals will be calculated by methods discussed in the previous chapter. That is to say that we estimate the quantities derived in chapter (3) with observed data by estimating $\hat{\alpha}_n$ for α, $\hat{\beta}_n$ for β, and $\mathrm{I}\{\mu_i(1, \hat{\beta}_n) \leq \mu_i(0, \hat{\beta}_n)\}$ for $g_{opt}(X_i)$. The only difference between the calculations in chapters 3 and what we will use here is due to the fact that the strategy g in chapter 3 is fixed where g is now β-dependent and this will need to be accounted for when computing the portion of the gradient matrix that uses numerical derivatives.

4.3 Simulation Studies

We report results of several simulations, each based on 10000 Monte Carlo data sets. We will use much of the same setup from chapter (2), we have data of the form $O_i = (D_i, E_i, X_i)$ where D_i is binary disease status, E_i is treatment, and (X_{1i}, X_{2i}) will be confounding covariates generated from a $N(0, 1)$ distribution independent of one another. For each given i^{th} patient we generated a Bernoulli treatment indicator E_i from the logistic regression model:

$$\text{logit}\{P(E_i = 1|X_i)\} = \alpha_0 + \alpha_1 X_{1i} + \alpha_2 X_{2i} + \alpha_3 X_{1i} X_{2i}, \tag{4.3}$$

and a Bernoulli disease indicator D_i from the logistic regression model:

$$\text{logit}\{P(D_i = 1|E_i, X_i)\} = \beta_0 + \beta_E E_i + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 E_i X_{1i} + \beta_4 E_i X_{2i}. \tag{4.4}$$

Once the data were generated, estimates of the model parameters were found for each Monte Carlo dataset using SAS logistic procedure (estimates and standard errors were output using the ODS delivery system). The generated data and logit model estimates were then input into SAS IML where estimates for AB_{opt}, standard errors, and confidence intervals were calculated for each Monte Carlo dataset. As in previous simulations we will use numeric derivatives in our standard error estimates that include a ε_j of 100^{th} of a standard error of each parameter estimate.

We will perform simulations in two broad categories; first we run simulations with one covariate to explore the doubly robust estimator of AB_{opt} under some of the scenarios put forth in chapter (2). In these first round of simulations we can note that

$\alpha_2 = \alpha_3 = \beta_2 = \beta_4 = 0$. In each case we have specified the correct model so we expect that the doubly robust estimator will not perform as well as a regression estimator. The second set of simulations will explore the two covariate case where will compare the regression estimator with the doubly robust estimator under a set of misspecified models.

4.3.1 One Covariate Simulations

Table 4.1 illustrates 3 examples with different combinations of model parameters, these examples are among the ones used to illustrate the effectiveness of the regression based estimator of AB_{opt}. We depict the true model parameters values along with estimates of $\hat{AB}_{opt}^{DR}$, Monte Carlo standard error, delta-theorem standard error and confidence intervals for both back-transformed and delta-theorem methods averaged over all 10000 Monte Carlo samples. We also include the median of the delta-theorem standard error of $\hat{AB}_{opt}^{DR}$, sample sizes of $n = 1000$ were used in all examples.

Examining table 4.1 we first notice the extreme skewness of $\hat{SE}(\hat{AB}_{opt}^{DR})_{dt}$, the average of the delta-theorem based standard error estimates of $\hat{AB}_{opt}^{DR}$, some of this is due to low propensity scores. Recall that we stated in chapter (3) that we needed to add an additional assumption that $0 < P(E = 1|X) < 1$ and if the $P(E = 1|X)$ becomes too small then inverse weights become large and estimators based on such methods can become biased. No restrictions were place on the data in table 4.1 in order to make fair comparisons between $\hat{AB}_{opt}^{DR}$ and $\hat{AB}_{opt}^{R}$. Considering the coverage probabilities for both the back-transform and delta-theorem based intervals, it is clear that only a few Monte Carlo datasets produced biased standard error estimates.

Instead of considering the average delta-theorem standard error estimate, let us look at the median standard error estimate of the Monte Carlo datasets. Comparing this to the average shows that the distribution of standard errors is indeed skewed, but note that the median delta-theorem standard error estimate is fairly close the Monte Carlo standard error of $\hat{AB}_{opt}^{DR}$ generated by finding the standard deviation of $\hat{AB}_{opt}^{DR}$ from each Monte Carlo simulation.

In comparing the quality of the doubly robust estimator to the model based estimator of AB_{opt} we will look at the asymptotic relative efficiency (ARE) [3] of one estimator to the other. Generally speaking the relative efficiency of one estimator to another is the

Table 4.1: Simulation Results; BT represent back-transformed confidence intervals while DT represents delta-theorem based confidence intervals

Simulation	1	2	3
n	1000	1000	1000
Reps	10000	10000	10000
α_0	-1.25	-1.25	-1
α_1	2	2	-1
β_0	-2.5	-2.5	1
β_1	1.25	1.25	1
β_2	1.1	1.1	1
β_3	0	2	-1
AB_{dom}	0.396	0.577	0.113
AB_{opt}	0.396	0.588	0.121
$\hat{AB}_{opt}^{DR}$	0.3983	0.5928	0.1239
$\hat{SE}(\hat{AB}_{opt}^{DR})_{mc}$	0.1129	0.0846	0.0238
$\hat{SE}(\hat{AB}_{opt}^{DR})_{dt}$	0.1012	0.2960	0.1627
Median$(\hat{AB}_{opt}^{DR})_{dt}$	0.0932	0.0777	0.0298
BT Lower 95%*	-1.3 x 10^{31}	-2.0 x 10^{62}	-3976.43
BT Upper 95%	0.5634	0.7792	0.3000
BT Coverage	0.9379	0.9502	0.9585
DT Lower 95%	0.2001	0.0127	-0.1949
DT Upper 95%	0.5966	1.173	0.4428
DT Coverage	0.9312	0.9413	0.9586
ARE	0.5008	0.4765	0.4983

*Due to skewed variance estimates, some back-transform confidence interval estimates are biased

ratio of their variances. Therefore we define

$$\text{ARE} = \left\{ \frac{\hat{SE}(\hat{AB}^{DR}_{opt})_{mc}}{\hat{SE}(\hat{AB}^{R}_{opt})_{mc}} \right\}^2$$

where $\hat{SE}(\hat{AB}^{R}_{opt})_{mc}$ can be found for each simulation in table 2.1. Recall that the simulations listed in table 4.1 are scenarios designed for the model based estimator to function at its best so it is no surprise to see a loss of efficiency in a doubly robust estimator when the model is correctly specified and low propensities could bias estimates. But even in these scenarios we learn that a doubly robust estimator for an estimated treatment strategy was close to the truth and confidence intervals had good coverage.

Recall the motivation for a doubly robust estimator came from an example where the model was not correctly specified. In shifting our focus to cases where the doubly robust estimator may be more appropriate we start by putting a bound on propensities. To keep inverse weights low we will restrict estimates of $P(E = 1|X)$ to fall between $(0.10, 0.90)$. Consider now an example where data is generated from the following propensity and outcome regression models:

$$\text{logit}\{P(E_i = 1|X_i)\} = -1.25 - X_i$$

and

$$\text{logit}\{P(D_i = 1|E_i, X_i)\} = -3 + 0.25E_i + 1.1X_i - 1.5E_iX_i.$$

Under this set of models $P(E = 1) = 0.3022$, $P(D = 1) = 0.0603$, $P(D^*(0) = 1) = 0.0451$, $P(D^*(1) = 1) = 0.0702$, and $P(D(g_{opt}) = 1) = 0.0343$.

Table 4.2: Simulation Results; BT represent back-transformed confidence intervals while DT represents delta-theorem based confidence intervals

Simulation	With Interaction	Without Interaction	With Interaction	Without Interaction	With Interaction	Without Interaction
n	2000	2000	2000	2000	2000	2000
Reps	10000	10000	10000	10000	10000	10000
AB_{dom}	0.4329	0.4329	0.4329	0.4329	0.4329	0.4329
AB_{opt}	0.2512	0.2512	0.2512	0.2512	0.2512	0.2512
	Regression Estimator		DR Estimator		Median of DR MC Samples	
$\hat{AB}_{opt}$	0.4390	0.1846	0.4445	0.2521	0.4466	0.2519
$\hat{SE}(\hat{AB}_{opt})_{mc}$	0.0760	0.0550	0.1219	0.0573		
$\hat{SE}(\hat{AB}_{opt})_{dt}$	0.0751	0.0550	0.6118	0.0572	0.1381	0.0569
BT Lower 95%	0.2698	0.0694	-1.16×10^{36}	0.1311	0.0150	0.1309
BT Upper 95%	0.5679	0.2855	0.7540	0.3561	0.7279	0.3561
BT Coverage	0.9396	0.7152	0.9617	0.9480		
DT Lower 95%	0.2918	0.0769	-0.7546	0.1401	0.0888	0.1398
DT Upper 95%	0.5862	0.2923	1.6436	0.3641	0.7745	0.3640
DT Coverage	0.9302	0.7541	0.9461	0.9469		

Listed in table 4.2 are the results of simulations where the data generated from the framework specified above and logit models were estimated with and without the interaction term in the outcome regression model. With the working model misspecified then estimates of AB will be trying to estimate AB_{dom} instead of the true AB_{opt}. As in earlier simulations we see that the regression based estimator does very well when the model is correctly specified but has poor coverage for both back-transform and delta-theorem based confidence intervals when the model is not correctly specified. When the interaction term is left out there is indeed bias in the estimation of the benefit of the dominant treatment.

By contrast we see the doubly robust estimator has excellent coverage for both intervals whether there was misspecification or not, even though AB estimates were estimating AB_{dom} and not AB_{opt}, the doubly robust estimator had better coverage for what it was intending to estimate. Again we see a skewness in the standard error estimates and medians are provided to show that a typical simulation run resulted in a delta-theorem standard error estimate that was close to the Monte Carlo standard error of $\hat{AB}^{DR}_{opt}$. We also see through the medians of the doubly robust estimator's confidence intervals that efficiency is still a potential problem because even with samples of size 2000 there is still a very wide confidence interval range.

4.3.2 Two Covariate Simulations

Returning to the general framework put forth in equations (4.3) and (4.4), we turn our attention to simulations involving more than confounding variable. The early simulations show that both estimators have good coverage for the true value of AB and we know that in cases where the correct model is specified that the model based estimator is far more efficient. The two covariate framework allows for a higher degree of model misspecification so that we can see what the effects of leaving out important variables can do to estimating attributable benefit. As such our approach will be similar to what was done in table 4.2, we will put forth one particular true model and study the effects of fitting an incorrect model to data generated from the truth. Consider now an example where data is generated from the following propensity and outcome regression models:

$$\text{logit}\{P(E_i = 1|X_i)\} = -2 + X_{1i} - 1.5X_{2i} + 0.5X_{1i}X_{2i}$$

and

$$\text{logit}\{P(D_i = 1|E_i, X_i)\} = -1 + 0.75E_i + X_{1i} + X_{2i} - E_iX_{1i} - E_iX_{2i}.$$

Under this set of models $P(E = 1) = 0.338$, $P(D = 1) = 0.358$, $P(D^*(0) = 1) = 0.305$, $P(D^*(1) = 1) = 0.438$, and $P(D(g_{opt}) = 1) = 0.249$.

Table 4.3 shows the results of several simulations of data fit from the framework proposed above. In the first simulation the correct model was specified estimating the AB_{opt}, as before we see better coverage and efficiency in the regression based estimator. In simulation 2 both the interaction terms were not included in the model, and thus the misspecified model is estimating AB_{dom}. Here we see better coverage from the doubly robust estimator, but the regression estimator's coverage was not as bad as in the one covariate scenario. So even though simulation 2 points to the doubly robust estimator being more accurate, the undercoverage is not very large and the regression estimates are not too bad. Note that the standard error estimates of the doubly robust and regression estimators are closer than in previous simulations. In simulation 3 we misspecify the propensity model by leaving out a mild interaction term which only affects doubly robust estimates, again we see the model based estimator dominate since the correct outcome regression model is specified, but the doubly robust estimator has good coverage.

Simulation 4 is an important one in that the misspecified model only leaves out one of the interaction terms in the outcome regression model, thus with one interaction term still in the outcome regression model the treatment regime whose benefit is being estimated is neither the optimal treatment strategy nor the one where everyone receives the dominant treatment. This is the first time we have encountered a treatment strategy whose benefit is neither AB_{opt} nor AB_{dom}, let's call this treatment strategy g and thus its benefit is AB_g. Note that the true value of AB_g is in between AB_{opt} and AB_{opt}, which in this setup is to be expected. The regression based estimate of AB_g is biased and has terrible coverage , coverage probabilities for back-transformed and delta-theorem based intervals are 0.1262 and 0.1358 respectively. By contrast we see the coverage probabilities for the same sets of intervals are 0.9440 and 0.9484 respectively, showing that doubly robust estimates well equipped to handle misspecification. In simulation 5 the misspecification involves both propensity and outcome regression models by leaving out all important interaction terms, for the regression based estimator this yields the same results as in simulation 2. In examining the doubly robust estimates from simulation 5 we still see better coverage from confidence intervals, although this is not guaranteed from the underlying assumptions.

Table 4.3: Simulation Results; BT represent back-transformed confidence intervals while DT represents delta-theorem based confidence intervals

Simulation	1	2	3	4	5
	Correct Model	Without β_3, β_4	Without α_3	Without β_4	Without $\alpha_3, \beta_3, \beta_4$
n	2000	2000	2000	2000	2000
Reps	10000	10000	10000	10000	10000
AB estimated	Optimal	Dominant	Optimal	Other	Dominant
AB true	0.3046	0.1468	0.3046	0.2688	0.1468
Regression Estimator					
$\hat{AB}_{opt}$	0.3067	0.1352	0.3067	0.2106	0.1352
$\hat{SE}(\hat{AB}_{opt})_{mc}$	0.0262	0.0231	0.0262	0.0188	0.0231
$\hat{SE}(\hat{AB}_{opt})_{dt}$	0.0258	0.0227	0.0258	0.0185	0.0227
BT Lower	0.2542	0.0896	0.2542	0.1735	0.0896
BT Upper	0.3555	0.1785	0.3555	0.2461	0.1785
BT Coverage	0.9461	0.9051	0.9461	0.1262	0.9051
DT Lower	0.2561	0.0908	0.2561	0.1744	0.0908
DT Upper	0.3573	0.1796	0.3573	0.2469	0.1796
DT Coverage	0.9446	0.9117	0.9446	0.1358	0.9117
Doubly Robust Estimator					
$\hat{AB}_{opt}$	0.3067	0.1455	0.3067	0.2579	0.1526
$\hat{SE}(\hat{AB}_{opt})_{mc}$	0.0361	0.0242	0.0400	0.0298	0.0260
$\hat{SE}(\hat{AB}_{opt})_{dt}$	0.1095	0.0238	0.1242	0.0727	0.0254
Median$\{\hat{SE}(\hat{AB}_{opt})_{dt}\}$	0.0371	0.0238	0.0440	0.0300	0.0250
BT Lower	-0.2596	0.0974	-14.13	0.0689	0.1012
BT Upper	0.4491	0.1910	0.4762	0.3739	0.2010
BT Coverage	0.9632	0.9457	0.9667	0.9440	0.9438
DT Lower	0.0921	0.0988	0.0632	0.1155	0.1027
DT Upper	0.5213	0.1922	0.5502	0.4003	0.2025
DT Coverage	0.9601	0.9464	0.9661	0.9484	0.9394

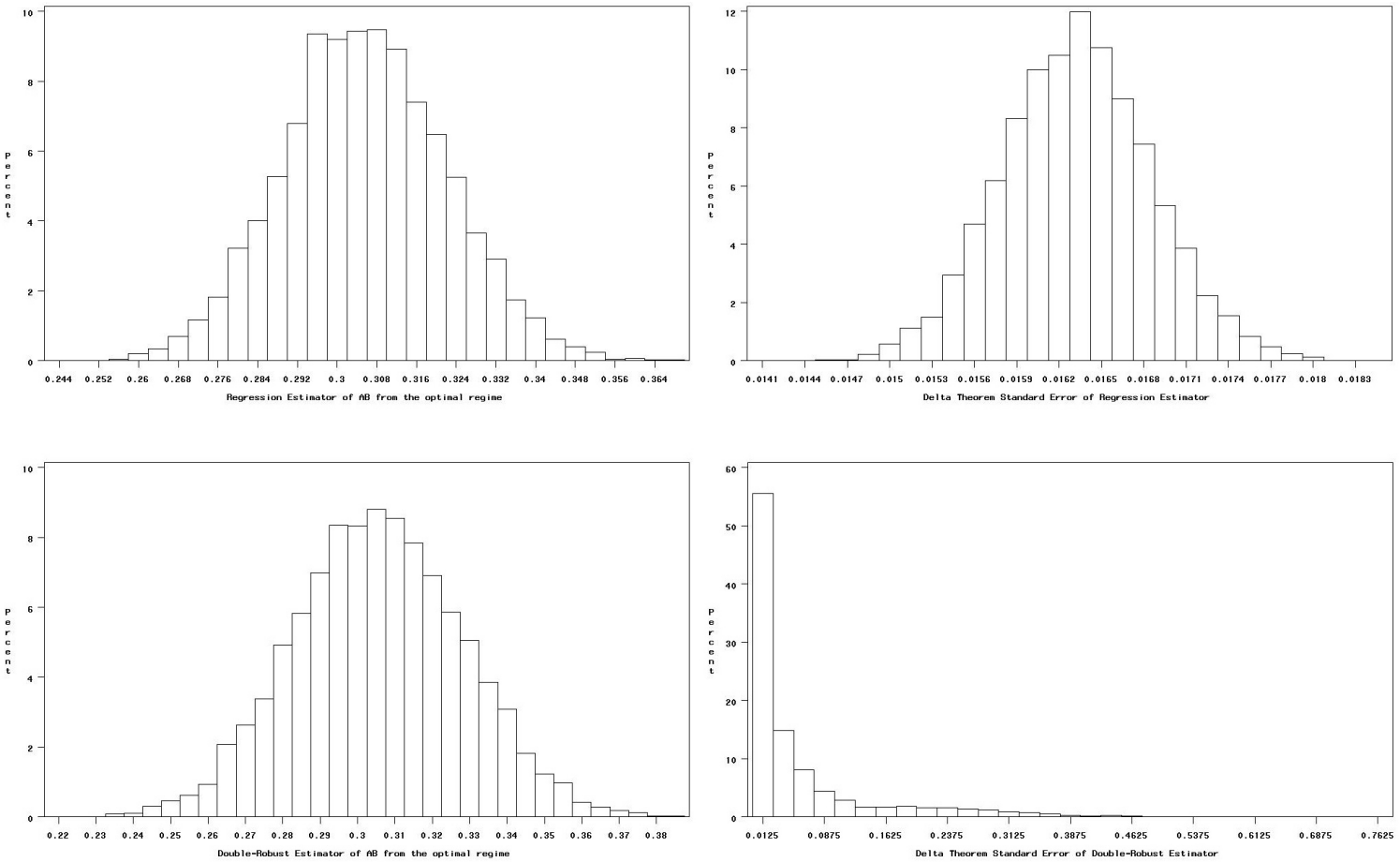

Figure 4.1: Histograms of both estimators and their standard errors under simulation 1 with sample size n=5000.

Overall we see a lot less variability in coverage in the doubly robust estimator in misspecified models than the regression estimator. There does seem to be a trade-off in efficiency and a skewness to variance estimates. To further explore this skewness, simulation 1 was repeated with a sample size of n= 5000. We see from the histograms listed in figure 4.1 that both the regression based and doubly robust estimators of AB are normally distributed, and while the regression based standard error is normal there is a definite skew to the doubly robust standard errors. A closer examination of the doubly robust estimator's standard errors can be found in figure 4.2. Even after heavily truncating the standard errors we still see a distinct skew to the data that not even a log transformation can balance out.

4.4 Example

We return to data collected from the Duke Databank for Cardiovascular Diseases (DDCD) to apply the doubly robust estimator to that dataset. Recall from chapter 2 that the DDCD contains nearly 40 years of baseline and follow-up data on patients undergoing cardiac catheterization. From this databank we focus on a subset of n=3856 patients who have already went through a prior cardiac artery bypass graft (CABG) and required later catheterization due to continued symptoms. We stated that for these patients the only treatment options were either medical therapy (MED) or a combination of medical therapy and percutaneous coronary intervention (PCI). Recall that the variable to indicate treatment assignment gave a value to 0 for those in the PCI group and 1 for those in the MED group, this was done so that the default treatment (in case $P(D = 1|E = 0, X) = P(D = 1|E = 1, X)$) would be the one with value 1. That is to say if patients are equally likely to die under both treatment strategies then that patient should receive medical therapy only.

Using several covariates we explored different models in table 2.2, we arrived at "best" model for this data was model 3 that had covariates congestive heart failure, number of diseased vessels, mitral insufficiency, pulse, and ejection fraction. Recall that 561 patients were missing a value for ejection fraction and 420 of those patients were in the PCI treatment group. Since an overwhelming majority of missing ejection fraction patients were given one particular treatment we concluded that the missing data was not missing at random and decided to add that data to our model by adding an indicator function of whether ejection fraction was missing or not.

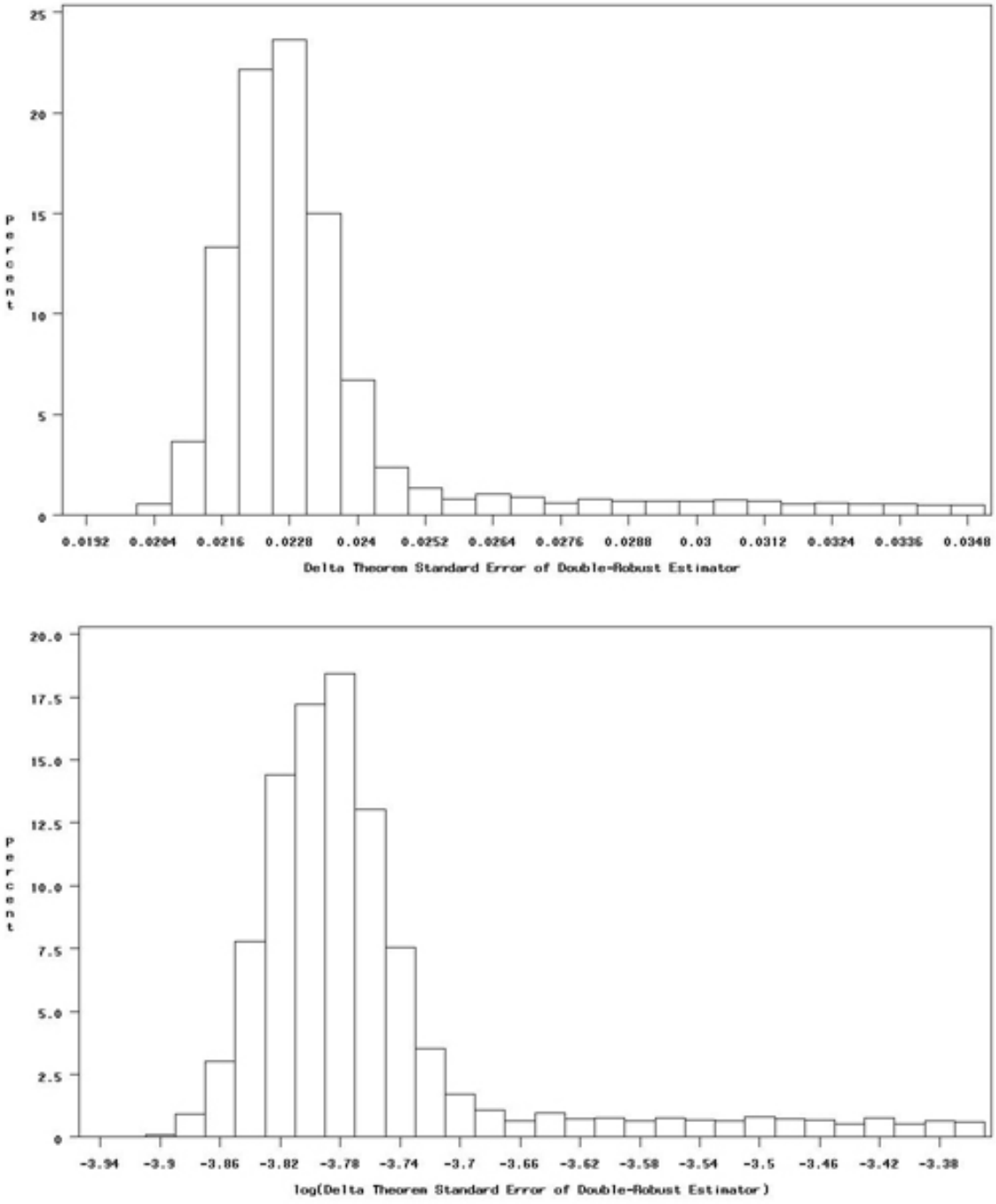

Figure 4.2: Histograms of delta-theorem doubly robust estimator standard errors and corresponding log transformed values, standard errors truncated at 0.035.

Table 4.4: Logistic Regression Models, analysis performed with SAS 9.1 logistic procedure.

Response: Death Within First Year	DF	Estimate	SE	Wald χ^2	P-value
Intercept	1	-4.8861	0.4385	124.16	$< .0001$
Treatment	1	0.3999	0.1601	6.24	0.0125
Congestive Heart Failure	1	0.6040	0.1420	18.10	$< .0001$
Number of Disease Vessels	1	0.3487	0.1013	11.86	0.0006
Mitral Insufficiency	1	0.2744	0.0619	19.64	$< .0001$
Pulse	1	0.0182	0.0042	18.81	$< .0001$
I(EJFX Present)	1	0.2215	0.2744	0.65	0.4194
Ejection Fraction*I(EJFX Present)	1	-0.0231	0.0049	22.08	$< .0001$

Response: Treatment (MED=1)	DF	Estimate	SE	Wald χ^2	P-value
Intercept	1	-1.9441	0.2445	63.23	$< .0001$
Congestive Heart Failure	1	0.5819	0.0853	46.52	$< .0001$
Number of Disease Vessels	1	-0.2088	0.0492	18.03	$< .0001$
Mitral Insufficiency	1	0.1558	0.0475	10.76	0.0010
Pulse	1	0.0133	0.0027	24.66	$< .0001$
I(EJFX Present)	1	2.3080	0.1821	160.61	$< .0001$
Ejection Fraction*I(EJFX Present)	1	-0.0024	0.0028	0.72	0.3950

In table 4.4 we see the parameter estimates and corresponding statistics for this constellation of covariates that form the "best" outcome regression model. In addition to the outcome regression model we also present the logistic regression analysis of a propensity model that determines treatment based on these same covariates. This model favors the optimal treatment regime where every patient is treated with a combination of angioplasty and medical therapy. Recall from earlier work with the regression based estimator of AB_{opt}, for this model $\hat{AB}^{R}_{opt} = 0.2157$ with confidence intervals $\{0.0356, 0.3621\}$ (back-transformed method) and $\{0.0536, 0.3777\}$ (delta method). The delta theorem based standard error of $\hat{AB}^{R}_{opt}$ is 0.0827. By contrast the doubly robust estimator of AB_{opt} yields an estimate of $\hat{AB}^{DR}_{opt} = 0.1741$ with confidence intervals $\{-0.0356, 0.3401\}$ (back-transformed method) and $\{-0.0118, 0.3599\}$ (delta method). The delta theorem based standard error of $\hat{AB}^{DR}_{opt}$ is 0.0948.

From these results we can draw several sharp contrasts between the estimators, most notable is the differences in the estimates themselves. We saw in simulation that in cases where the regression model was specified correctly, both estimates were close to the

Table 4.5: Logistic Regression Models, analysis performed with SAS 9.1 logistic procedure.

Response: Death Within First Year	DF	Estimate	SE	Wald χ^2	P-value
Intercept	1	-5.0792	0.4564	123.86	$< .0001$
Treatment	1	0.3581	0.1598	5.02	0.0250
Congestive Heart Failure	1	1.1714	0.2972	15.53	$< .0001$
Number of Disease Vessels	1	0.3287	0.1020	10.38	0.0013
Mitral Insufficiency	1	0.2742	0.0622	19.42	$< .0001$
Pulse	1	0.0178	0.0042	17.87	$< .0001$
I(EJFX Present)	1	0.7036	0.3562	3.90	0.0482
Ejection Fraction*I(EJFX Present)	1	-0.0231	0.0050	24.96	$< .0001$
I(EJFX Present)*CHF	1	-0.7305	0.3340	4.78	0.0287

Response: Treatment (MED=1)	DF	Estimate	SE	Wald χ^2	P-value
Intercept	1	-2.2025	0.2566	73.69	$< .0001$
Congestive Heart Failure	1	1.5436	0.2136	52.20	$< .0001$
Number of Disease Vessels	1	-0.2332	0.0503	21.53	$< .0001$
Mitral Insufficiency	1	0.1485	0.0472	9.91	0.0016
Pulse	1	0.0126	0.0027	21.94	$< .0001$
I(EJFX Present)	1	2.8264	0.2151	172.61	$< .0001$
Ejection Fraction*I(EJFX Present)	1	-0.0043	0.0029	2.25	0.1340
I(EJFX Present)*CHF	1	-1.1385	0.2310	24.29	$< .0001$

"true" value. For this problem we see a difference in AB_{opt} estimates of 0.0416, a substantial difference given both estimators are supposed to be unbiased if the true model is correctly specified and one considers that the standard errors under both estimation techniques were between 0.08 and 0.10. The doubly robust estimators intervals suggest that the true AB_{opt} may indeed be zero, where as the regression estimators intervals suggest significant benefit in treating everyone with medical therapy and angioplasty (PCI). These differences lead one to conclude that the outcome regression model may be misspecified and some important terms were left out. Recall that standard model selection techniques (backward and forward variable selection) were employed in our search for the "best" model. Given the large number of covariates that were potentially confounding variables, an assumption was made to include only main effects and interactions between treatment and significant main effects in our selection process. That is to say that model selection techniques were restricted to main effects and then those main effects were tested for interactions with treatment.

Expanding the search for confounding covariates we find at least one more con-

founding variable in the form of an interaction between the indicator of ejection fraction and whether the patient has experienced congestive heart failure. Examining table 4.5 we have added this new confounder to our models and form new estimates of AB_{opt} with $AB^{R}_{opt} = 0.1957$ with delta theorem standard error 0.0832 which yields a confidence interval of $\{0.0325, 0.3588\}$. From doubly robust estimation we have $AB^{DR}_{opt} = 0.1686$ with delta theorem standard error 0.0940 which yields a confidence interval of $\{-0.0157, 0.3529\}$.

After examining both the output from both estimators we see that using doubly robust techniques provided valuable insight into this problem. As an estimator of attributable benefit it seemed to perform well in real life data analysis, for this case we see that adding one significant covariate to our models dropped the regression based estimator by 0.02 but the corresponding doubly robust estimator fell by only 0.0055. Even though no formal arguments have been made to determine if the doubly robust estimator performs better than the model based estimator; in simulation and real data where confounders were left out of the propensity and outcome regression models we have seen better performance and less sensitivity to adjustments in estimated models from the doubly robust estimator. Lastly the reader should recall that to reduce bias in inverse propensity weights, an assumption was made to keep $0 < P(E = 1|X) < 1$; the propensities from models formed with this data all fell between (0.10, 0.93) therefore no data were excluded from the analysis due to low propensity scores.

Chapter 5

Discussion

In this thesis we have introduced the notion of attributable benefit associated with a treatment strategy. This measure can be of public health interest and represents the fraction of events that could have been prevented by using a particular treatment regime. Similar in spirit to adjusted attributable risk, but has a very different interpretation and can handle a wider variety of treatment strategies. This measure is most useful in the case where one treatment may not be superior to another for all patients.

We have developed two estimators of attributable benefit, the first uses an outcome regression model to estimate the proportion of disease cases that would have occurred had a particular strategy been used (referred to as $P\{D^*(g) = 1\}$). The second estimator uses propensity scores models and outcome regression models to form an estimate of $P\{D^*(g) = 1\}$ that provides doubly robust protection from model misspecification. Notions from statistical causality were used to lay the ground work for using both estimators and standard errors were calculated using large sample theory through the use of influence functions. Two types of confidence intervals were obtained for both estimators; the back-transformed interval and the delta method interval.

Simulation studies were performed to examine both the regression based and doubly robust estimators of attributable benefit under a variety of scenarios. In comparing both estimators we find that while each estimator has its own merits, no definitive conclusion can be made as to whether one estimator is better than the other. We found through simulation that the regression based estimator, $\hat{AB}_g^R$, was a more efficient estimator than the doubly robust estimator, $\hat{AB}_g^{DR}$ when the model for outcome regression was correctly specified. When the correct outcome regression model was specified then both estimators had good

coverage probabilities, but $\hat{AB}_g^R$ became unstable and coverage is poor if important terms are left out of the outcome regression model.

In addition, both estimators were applied to a real-life dataset from the Duke Databank for Cardiovascular Disease in an attempt to answer a question of clinical importance. Standard model selection techniques (backward and forward variable selection) were employed to find the "best" outcome regression model for the data. Early in the analysis it was decided to use only main effects of confounders and interactions between treatment and confounders with significant main effects as possible covariates in the outcome regression model. Confidence intervals of AB_{opt} from $\hat{AB}_{opt}^R$ suggested that there was significant benefit to employing a strategy that favored using the dominant treatment (a combination of medical therapy and angioplasty) for all patients. Later estimates and confidence intervals of AB_{opt} from $\hat{AB}_{opt}^{DR}$ suggested that the outcome regression model was misspecified and a significant interaction term between two confounders was added to the model. In the end although the two estimators gave results in the same direction and estimates that were within a half a standard error of each other, the attributable benefit was not statistically significant based on the doubly robust estimator at the 0.05 level whereas the outcome regression estimator was statistically significant.

Since neither method yielded a superior estimator of attributable benefit, perhaps there is merit in calculating both $\hat{AB}_g^R$ and $\hat{AB}_g^{DR}$ and using agreement between the estimates as a diagnostic indicator of whether the outcome regression model is correctly specified. This diagnostic indicator may be better suited in the form of z-scores, from the asymptotic theory developed earlier we can note

$$Z_1 = \frac{\hat{AB}_g^R}{SE\{\hat{AB}_g^R\}}$$

and

$$Z_2 = \frac{\hat{AB}_g^{DR}}{SE\{\hat{AB}_g^{DR}\}}$$

follow a $N(0,1)$ distribution. From the DDCD example the outcome regression model without the interaction (I(EJFX Present)*CHF) term, $Z_1 = 2.61$ and $Z_2 = 1.84$ (using delta-theorem standard errors). While no formal analysis has been put forth to study the relationship between Z_1 and Z_2, it is obvious in this case that there was a discrepancy between the two estimators.

5.1 Limitations and Future Work

We acknowledge a difficulty of the asymptotic theory under the null hypothesis of no treatment effect, that is $P(D = 1|E = 1, X) = P(D = 1|E = 0, X)$ for all X. Though if this null hypothesis were true then estimating AB may not be of interest. We suggest that this methodology be used only after the data analyst is comfortable that there is a significant treatment effect. We have not studied the statistical properties of such a two-stage approach, where at the first stage a formal test is conducted to establish whether there is evidence of a treatment difference, and only if there is significant treatment difference do we consider estimating attributable benefit for an optimal treatment regime. This would be interesting for future research.

The cases discussed so far involved only two treatments but the methodology can be extended to many treatment cases. In the case where there are r treatments then we define $h(X_i, \beta) = \sum_{j=1}^{r} \mu_i(j, \beta)\mathrm{I}[\mu_i(j, \beta) = \min_{0 \le r_2 \le j} \mu_i(r_2, \beta)]\mathrm{I}[\mu_i(j, \beta) < \min_{j < r_2 < r} \mu_i(r_2, \beta)]$. The corresponding optimal treatment regime would assign treatment based on which treatment has the smallest value of $\mu_i(j, \beta)$ and if many treatments have the same value of $\mu_i(j, \beta)$ then the largest numbered treatment is assigned. The outcome regression estimates of AB_{opt} and confidence intervals follow exactly as discussed earlier, though it is unclear whether doubly robust estimates will also follow in the same manner.

Calculations of standard errors of estimators or AB_g can be quite tedious, and should only be attempted with the use of computer software. A SAS macro has been developed that will produce estimates and standard errors of AB^R_{opt} for any logistic regression model that can be specified in the SAS logistic procedure. The code for this macro has been added as an appendix to this thesis, a macro is being developed that gives estimates and standard errors of AB^{DR}_{opt}, but at the time of this publication is not ready for general use. We acknowledge that the widespread use of such methods will not be achieved until adequate software is developed and available for public use.

Lastly, no framework for hypothesis testing has been given in this manuscript. Can we test whether there is significantly more attributable benefit from one regime to another (both based on the same model)? For example say we have the correct statistical model and one calculates AB_{opt} and AB_{dom}. One may ask whether there is significantly more benefit from using a treatment regime that adjusts for patients according to their covariates versus

a more static regime where everyone gets the better treatment. In addressing that question we believe the measure of interest is $E\{D^*(g_{opt})\} - E\{D^*(g_{dom})\}$. This is also potential avenue of future work.

Bibliography

[1] Al-Khatib, S.M., Shaw, L.K., O'Connor, C., Kong, M., and Califf, R.M. (2007) Incidence and Predictors of Sudden Cardiac Death in Patients with Diastolic Heart Failure. *Journal of Cardiovascular Electrophysiology* **18,** 1231-1235.

[2] Benichou, J. (2001) A review of adjusted estimators of attributable risk. *Statistical Methods in Medical Research* **10,** 195-216.

[3] Casella, G. and Berger, R.L. (2002) *Statistical Inference, Second Edition.* Duxbury Press, Belmont, CA.

[4] Graubard, B.I. and Fears, T.R. (2005). Standard Errors for Attributable Risk for Simple and Complex Sample Designs. *Biometrics* **61,** 847-855.

[5] Jewell, N.P. (2004) *Statistics for Epidemiology.* Chapman & Hall, London, United Kingdom.

[6] Levin, M.L. (1953) The occurrence of lung cancer in man. *Acta Unio Internationalis Contra Cancrum* **9,** 531-541.

[7] Neyman, J. (1923) Sur les applications de la thar des probabilities aux experiences Agaricales: Essay de principle. English translation of excerpts by Dabrowska, D. and Speed, T. (1990). *Statistical Science* **5,** 465-480.

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

[9] Scharfstein, D.O., Rotnitzky, A., and Robins, J.M. (1999) Adjusting for nonignorable drop-out using semiparametric nonresponse models (with discussion). *Journal of the American Statistical Association* **94,** 1096-1146.

[10] Stefanski, L. and Boos, D. (2002) The Calculus of M-Estimation. *The American Statistician* **56,** 29-38.

[11] Tsiatis, A.A. (2006) *Semiparametric Theory and Missing Data.* Springer, New York.

[12] Whittemore, A.S. (1982) Statistical methods for estimating attributable risk from retrospective data. *Statistics and Medicine* **1,** 229-243.

Appendices

```
/*****************************************************************
|
| Program Name:  AB opt covariate simulation
|
| Program Version: 1.0
|
| Program Purpose:  Calculates AB for the optimal treatment regime from a
| prespecified model.  User must preload data into SAS work folder.
|
| SAS Version: 8 or 9
|
| Created By:   Jason Brinkley
| Date:            27-Mar-2008
|
|*****************************************************************
| Change Log
|
| Modified By:
| Date of Modification:
|
| Modification ID:
| Reason For Modification:
|
*****************************************************************/

/*Instructions:
Load data into SAS work folder. Data must be of form: Binary Disease,
Binary Exposure/Treatment, Continuous or Binary Covariates.  For discrete
variable input, analyst needs to make binary indicators for each level
of covariate and input those binary indicators instead of discrete
variables.User must create interaction variables in the dataset before
running macro.  Analyst needs to use model selection techniques to find
"best" model before running macro.

Macro inputs are as follows:

alpha - Confidence level for intervals
(i.e. 95% intervals mean alpha=0.05)

Data - name of dataset in work folder

D -  Binary outcome/disease variable
```

```
E -  Binary exposure/treatment variable

X -  one or many covariates.  Continuous or Binary only.
Interactions between covariates go here.

X2 -  Which covariates/exposure interactions are included in the model
(i.e. if E*X is significant then put X here)

Int -  User created interaction variables in model
(i.e. if E*X is significant, analyst creates variable EX=E*X
and it goes here)

Out -  Name of output dataset for further manipulation

Interaction -  Indicator of whether there is covariate/exposure
interactions in the model (0 = no interactions, 1 = interactions).
IF THERE ARE COVARIATE/COVARIATE INTERACTIONS ONLY THEN PUT 0 HERE.

*/

%Macro AB_opt_reg(alpha, Data, D, E, X, X2, Int, Out, Interaction);

*Different logit models whether there is interactions;

%IF &Interaction = 0 %Then %Do;
proc logistic descending data=&Data;
model &D = &E &X;
ods output ParameterEstimates = ParameterEstimates;
run;
%End;

%IF &Interaction = 1 %Then %Do;
proc logistic descending data=&Data;
model &D = &E &X &Int ;
ods output ParameterEstimates = ParameterEstimates;
run;
%End;

quit;

*Betas from logit models to be used in AB analysis;
Data Betas;
set ParameterEstimates;
Keep Data Variable Estimate StdErr;
```

```
run;


*Output file in html with given filename;
ods html body=&Output_file;

Proc iml;

*************************************************************************
*   influence function for AB is calculated in 4 parts
*   using P_Dstar0, IF1, IF2, IF3;
*
*************************************************************************;

               *Output files step, initializing variables;
               P_Dopt=0;
               P_D=0;
               Lower_BT=0;
               Upper_BT=0;
               AB_opt_hat = 0;
               ln_ratio = 0;
               lower_DT =0;
               upper_DT = 0;
               se_ab=0;

                     *Load data;
                     use sample;
                     read all var {&X} into X;
                     read all var {&E} into E;
                     read all var {&D} into D;
                     read all var {&X2} into X2;

if &Interaction = 1 then read all var {&Int} into int;


*Calculate estimates for P(D=1);
P_D=(sum(D))/(nrow(D));

*IF3 calculation;
IF3=(D-P_D)/P_D;

                     Use Betas;
                     read all var{estimate} into B;
                     read all var{stderr} into StdErrB;
```

```
*bound needed for numeric derivatives;
bound = .01 * StdErrB;
bound2= bound;
r = nrow(B);
n=nrow(D);

Outvar = n||P_Dopt || P_D || AB_opt_hat ||ln_ratio || Lower_BT ||Upper_BT
||Lower_DT || Upper_DT ||SE_AB  ;

cname = {"Sample Size" "Prob D optimal" "Prob Disease" "AB opt hat"
"ln(ratio)" "Lower BT" "Upper BT" "Lower DT" "Upper DT" "Delta SE" };

create out from Outvar [ colname=cname ];


*f, f_0, and f_1 are different whether there are interactions or not;
I=j(n,1);
if &Interaction=0 then f=T(I||E||X);
if &Interaction=1 then f=T(I||E||X||Int);

I1=j(n,1);
*I2 is a vector of zeros;
I2=I1-I1;
if &Interaction=0 then f_0=T(I1||I2||X);
if &Interaction=1 then f_0=T(I1||I2||X||I2);

if &Interaction=0 then f_1 = T(I1||I1||X);
if &Interaction=1 then f_1 = T(I1||I1||X||X2);

E_opt = E;

*create E_opt, E_opt chooses level with lowest chance of poor outcome.
If the chance of a poor outcome is the same then macro defaults to trt 1;

do k = 1 to n;
E_opt[k,] = 0;
M1 = exp(T(B)*f_0[,k]);
M2 = 1/(1+exp(T(B)*f_0[,k]));
M=M1*M2;

W1 = exp(T(B)*f_1[,k]);
W2 = 1/(1+exp(T(B)*f_1[,k]));
W=W1*W2;
```

```
if W < M then E_opt[k,]=1;
end;


*X2 may be a subset of X variables;

if &Interaction=1 then EoptX = E_opt#X2;

if &Interaction=0 then f_opt = T(I1||E_opt||X);
if &Interaction=1 then f_opt = T(I1||E_opt||X||EoptX);

avg1 =0;

*P_Dopt calculation and IF1 calculation;

M1 = exp(T(B)*f_opt);
M2 = 1/(1+exp(T(B)*f_opt));
M=M1#M2;
avg1=M[,:];

IF1=M-avg1;
IF1=T(IF1);

P_Dopt=avg1;

*IF2 Calculation;

Avg2=0;

ncol = nrow(f);
Temp = I(ncol);

*avg is a matrix of zeros;
Avg3=T(Temp - Temp);

*numeric derivatives;
row_vector = j(1,r,1);
LowB = B*row_vector;
UpB = LowB;

*need these for the numeric derivatives later;
do v = 1 to r;
LowB[v,v] = B[v,]-bound[v,];
```

```
UpB[v,v] = B[v,]+bound[v,];
bound2[v,]=.5*(1/bound[v,]);
end;

*numeric Derivative loop;
ND1 = B;
ND2 = B;

Do v = 1 to r;
NM1 = exp(T(LowB[,v])*f_opt);
NM2 = 1/(1+exp(T(LowB[,v])*f_opt));
NM3 = sum(NM1#NM2)/n;
ND1[v,1] = NM3;

NM4 = exp(T(UpB[,v])*f_opt);
NM5 = 1/(1+exp(T(UpB[,v])*f_opt));
NM6=sum(NM4#NM5)/n;

ND2[v,1]= NM6;
end;


*ND is numeric approximation for partial mu/partial beta;
ND = (ND2 - ND1)#bound2;


Do k = 1 to n;
Q1=f[,k]*T(f[,k]);
Q2 = exp(T(B)*f[,k]);
Q3 = (1+exp(T(B)*f[,k]))*(1+exp(T(B)*f[,k]));
Q3=1/Q3;
Q4=(Q2*Q3)*Q1;
Avg3 = Avg3 + Q4;
end;

Avg3 = Avg3/n;
Avg3 = inv(Avg3);
IF2 = T(ND)*Avg3*f;
IF2 = T(IF2);

R1 = exp(T(B)*f);
R2 = 1/(1+exp(T(B)*f));
R=T(R1#R2);
IF2 = IF2 # (D-R);
```

```
    *Put all the pieces together to get IF;
    IF = ((IF1+IF2)/p_dopt)-IF3;
    V = T(IF) * IF;
    V = V / ((n-1)*(n-1));
    V = sqrt(V);

*V is estimate of the standard error of ln(PDstar0=1)-ln(P D=1);

    Ratio=P_Dopt/P_D;
    AB_opt_hat = 1-ratio;
    ln_ratio =log(ratio);

*Confidence Intervals;
alpha=&alpha;
z = probit(1-(alpha/2));

*Back Transform 95% Confidence Interval;
Lower_BT =1-exp(ln_ratio + (z* V));
Upper_BT=1-exp(ln_ratio - (z*V));

*Delta Theorem 95% Confidence Interval;
Lower_DT = AB_opt_hat - (z*ratio*V);
Upper_DT = AB_opt_hat + (z*ratio*V);
SE_AB = ratio*V;

*Output needed values;
Outvar = n || P_Dopt || P_D || AB_opt_hat ||ln_ratio || Lower_BT ||
Upper_BT||Lower_DT || Upper_DT ||SE_AB ;
append from Outvar;

run;


**********************************************************************
*
* Start of analysis for output
*
**********************************************************************;

Data &out;
set out;
                label
                Sample_Size = 'Sample Size'
```

```
                ln_ratio_ = 'Log{P(D)/P(D_opt)}'
                Prob_D_optimal = 'Probability D optimal'
                Prob_Disease = 'Probability of Disease'
                AB_opt_hat = 'AB opt hat'
               Lower_BT = 'Backtransformed Lower 95% C.I.'
               Upper_BT = 'Backtransformed Upper 95% C.I.'
               Lower_DT = 'Delta Thm Lower 95% C.I.'
               Upper_DT = 'Delta Thm Upper 95% C.I.'
Delta_SE = 'Delta Thm Standard Error';
run;

Proc Print data=Out label;
run;

quit;
quit;

%MEND;
```