Peer Reviewed

Title:
The Relative Performance of Targeted Maximum Likelihood Estimators Under Violations of the Positivity Assumption

Author:
Porter, Kristin Elizabeth

Acceptance Date:
2011

Series:
UC Berkeley Electronic Theses and Dissertations

Degree:
Ph.D., BiostatisticsUC Berkeley

Advisor(s):
van der Laan, Mark J.

Committee:
Hubbard, Alan, Sekhon, Jasjeet S.

Permalink:
http://escholarship.org/uc/item/3hp4r33n

Abstract:

The Relative Performance of Targeted Maximum Likelihood Estimators Under Violations of the Positivity Assumption

by

Kristin Elizabeth Porter

A dissertation submitted in partial satisfaction of the
requirements for the degree of
Doctor of Philosophy

in

Biostatistics

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Mark J. van der Laan, Chair
Professor Alan Hubbard
Professor Jasjeet S. Sekhon

Spring 2011

The Relative Performance of Targeted Maximum Likelihood Estimators Under Violations of the Positivity Assumption

Abstract

The Relative Performance of Targeted Maximum Likelihood Estimators Under Violations of the Positivity Assumption

by

Kristin Elizabeth Porter

Doctor of Philosophy in Biostatistics

University of California, Berkeley

Professor Mark J. van der Laan, Chair

Observational studies often present the challenge of data sparsity due to violations of the positivity assumption. Such violations occur when some subgroups never or rarely receive a particular treatment or never or rarely are uncensored. Bias due to actual or practical positivity violations often goes undiagnosed, and such bias can threaten valid inference for estimation of the target parameter. It is important to recognize that different estimators perform differently under a lack of positivity - in terms of both bias and variance. Common estimators across many fields often perform poorly in this setting.

Alternatively, targeted maximum likelihood estimators (TMLE's) tend to be relatively robust under a lack of positivity. This dissertation compares the performance of TMLE's to many common estimators under violations of the positivity assumption for three different target parameters: (1) a causal effect focused on the difference in mean outcomes for two treatments, (2) a mean outcome that is subject to missingness but for which all possible covariates for predicting missingness are measured, and (3) conditional relative risk in a semi-parametric multiplicative regression model.

For each of these parameters, the parameter-specific positivity assumption is formally presented and discussed. Also for each parameter, the theoretical properties of existing methods are compared to the those of TMLE's. The theoretical properties indicate how we expect different estimators to behave under positivity violations. Also, using a variety of simulations with various degrees of and reasons for positivity violations, the performance of TMLE's, relative to other estimators, is demonstrated. This dissertation also discusses how to diagnose bias due to positivity violations and how to respond to resulting bias.

Contents

List of Figures

List of Tables

Acknowledgments

I am extremely thankful to my PhD advisor, Mark J. van der Laan, for his guidance, mentorship and support over the past three years. I also thank many individuals who contributed to the work presented in this thesis, including Maya Petersen, Jasjeet S. Sekhon, Susan Gruber, Yue Wang and Catherine Tuglus. I also am appreciative of Alan Hubbard for his guidance and support as my MA advisor. Finally, I would like to thank Mark J. van der Laan, Alan Hubbard and Jasjeet S. Sekhon for their review of this dissertation and their helpful comments.

Chapter 1

Introduction

A rigorous, statistical investigation of a scientific question involves not only a good research design, reliable data, and an appropriate target parameter, but also careful consideration of estimation methods for the target parameter. Analysts too often use standard estimation methods, simply relying on status-quo procedures, popular software packages and unvalidated assumptions. However, there are many choices to be made when considering different estimation methods, which can affect the bias and efficiency of results. While it may seem obvious to state, it is not always fully appreciated that estimation methods matter, often considerably. This dissertation illustrates this point.

What choices must an analyst weigh when making decisions about estimation methods? One key choice is the *estimator* of the target parameter. For example, one may choose an estimator that relies on a model for the conditional expectation of the outcome given covariates; or, one may choose an estimator that relies on a model for the conditional expectation of treatment or censoring given covariates (i.e. propensity). Alternatively, an analyst may choose an estimator that incorporates both of these models - i.e. a double robust (DR) estimator. For each type of estimator, there are a multitude of estimators to consider.

Another key choice when deciding on an estimation method, is the *statistical model* on which a candidate estimator relies. For example, an analyst may use parametric models, which rely on substantive knowledge to select covariates and a functional form. In contrast, one may rely on non-parametric models, which instead let the data speak through data-adaptive, or machine learning, algorithms, or one may opt for a semi-parametric model, which combines both parametric assumptions and non-parametric, data-adaptive methods.

When weighing these choices, an analyst should consider many factors. For example, the research design is one important factor. If one has a successful randomized control trial (RCT), the choices of estimator and corresponding statistical model may not affect consistency, but they can be very important in terms of efficiency (Rosenblum and van der Laan [2010], Moore and van der Laan [2007]). On the other hand, in observational studies, these choices can have a substantial effect on both consistency and efficiency, as will be demonstrated throughout this dissertation. In both types of studies, many other factors can

also affect the performance of different estimation methods, including data structure (such as a nested or longitudinal structure vs. a point-treatment structure), censoring, and of course, confidence in model assumptions.

One other key factor that can have substantial implications for choices related to the estimation method, particularly related to the choice of estimator, is whether or not there is a "lack of positivity" in the data. Lack of positivity, a common challenge in observational data, occurs when there is a lack of support in the data for some subgroup(s) of subjects because they never or rarely receive some treatments or never or rarely are not censored. More formally, lack of positivity arises when there are either actual or theoretical (i.e. practical) violations of what is referred to as the positivity assumption (see Robins [1986, 1987a, 1999], Petersen et al. [2010]) or the experimental treatment assignment (ETA) assumption (Neugebauer and van der Laan [2005]). The assumption is parameter specific. For a causal effect, it states that within each stratum of covariates, there is a positive probability for all possible treatment assignments. For a mean outcome under missingness, it requires that within each stratum of covariates, there is positive probability that the outcome is not missing. Identifiability of the target parameter requires that the appropriate positivity assumption is not violated. However, even if the assumption holds, theoretical violations can lead to substantial bias, with or without inflated variance. The extent of bias and/or variance inflation depends greatly on the estimator. Some estimators are constructed in a such a way that they are much more robust to positivity violations.

Targeted maximum likelihood estimators (TMLE's) make up a class of estimators that are relatively robust to positivity violations compared to many other estimators that are either used widely or are found in methodological literature across many disciplines. There is a growing set of literature on TMLE's (for a summary, see van der Laan et al. [2009] and Rose and van der Laan [Eds.]), which exist for any parameter of interest, including those defined by non-parametric, semi-parametric and parametric models. TMLE's are double robust and asymptotically efficient, substitution estimators that are obtained by fluctuating an original estimate of the density of the data in a way that targets the parameter of interest. Because they are substitution estimators, TMLE's respect the fact that the true parameter value is a particular function of the data generating probability distribution in the assumed statistical model. Because of this property, TMLE's respect the global constraints on the data generating distribution and the parameter space. This results in their relative robustness under lack of positivity.

TMLE's can also incorporate data-adaptive likelihood or loss-based estimation procedures to estimate both the conditional expectation of the outcome and of the missingness or treatment mechanism. Moreover, TMLE's allow for the incorporation of targeted estimation of the censoring/treatment mechanism, which is introduced in what is referred to as the collaborative TMLE (C-TMLE), thereby fully confronting a long-standing problem of how to select covariates in the missingness or treatment mechanism of DR estimators. Such an extension results in even greater improvements in robustness in the face of positivity violations.

In summary, this dissertation makes a clear point that choices related to estimation methods can really matter, particularly in observational studies. Observational studies often present the challenge of data sparsity due to a lack of positivity. Bias due to a lack of positivity often goes undiagnosed. Such bias can threaten valid inference for estimation of the target parameter. It is important to recognize that different estimators perform differently under lack of positivity - in terms of both bias and variance. Common estimators across many fields often perform poorly in this setting. Alternatively, TMLE's tend to be relatively robust under lack of positivity.

This dissertation compares the performance of many common estimators to TMLE's under violations of the positivity assumption for three different target parameters: (1) a causal effect focused on the difference in mean outcomes for two treatments, (2) a mean outcome that is subject to missingness but in which all possible covariates for predicting missingness are measured, and (3) conditional relative risk in a semi-parametric multiplicative regression model. For all parameters, the parameter-specific positivity assumption is formally presented and discussed. Also, for all parameters, the theoretical properties of existing methods are compared to the theoretical properties of TMLE's. The theoretical properties indicate how one expects the different estimators to behave under lack of positivity. Then, the relative performance of the estimators is demonstrated through a variety of simulations with various degrees of and reasons for positivity violations. The dissertation also discusses how to diagnose bias to positivity violations, and how to respond to resulting bias.

To illustrate these points, the chapters of this dissertation are summarized as follows:

- Chapter 1 discusses the positivity assumption in the context of assessing model and parameter-specific identifiability of causal effects. In this case, positivity violations occur when certain subgroups in a sample rarely or never receive some treatments of interest. Also, the parametric bootstrap is presented as a tool to assess the severity of threats to valid inference due to positivity, and its utility as a diagnostic is explored using simulated data. Several approaches for improving the identifiability of parameters in the presence of positivity violations are also reviewed. All of the approaches can be understood as trading off proximity to the initial target of inference for identifiability, and should be considered systematically. This chapter is closely based on the published technical reported titled "Diagnosing and Responding to Violations in the Positivity Assumption" by Maya L. Petersen, Kristin E. Porter, Susan Gruber, Yue Wang, Mark J. van der Laan (Petersen et al. [2010]).

- Chapter 2 delves more deeply into the relative performance of TMLE's under lack of positivity, while focusing on a simple missing data problem in which one wishes to estimate the mean of an outcome that is subject to missingness and covariates predicting missingness are measured. Based on a draft of an article titled "The Relative Performance of Targeted Maximum Likelihood Estimators" by Kristin E. Porter, Susan Gruber (co-first authors), Mark van der laan and Jasjeet S. Sekhon, this chapter highlights an active debate in the literature on censored data about the relative

performance of model based maximum likelihood estimators, inverse probability of weighting (IPCW) estimators, and a variety of DR, semi-parametric efficient estimators. In particular, Kang and Schafer [2007] demonstrate the fragility of DR and IPCW estimators in a simulation study with positivity violations. Responses by Robins et al. [2007], Tsiatis and Davidian [2007], Tan [2007] and Ridgeway and McCaffrey [2007] further explore the challenges faced by double robust estimators and offer suggestions for improving their stability. In this chapter/article, we join the debate by presenting several TMLE's. We explain that TMLE's, particularly those that guarantee that the parametric submodel employed by the TMLE procedure respects the global bounds on the continuous outcomes, are especially suitable for dealing with positivity violations because in addition to being DR and semi-parametric efficient, they are substitution estimators. We also demonstrate the practical performance of TMLE's relative to other estimators in the simulations designed by Kang and Schafer [2007] and in modified simulations with even greater estimation challenges.

- Chapter 3 focuses on a parameter that is defined by a semi-parametric model. Specifically, it introduces two TMLE's of conditional relative risk in a semi-parametric multiplicative regression model for a binary outcome. The introduction of the semi-parametric model is an approach that responds to bias due to positivity violations but that maximizes flexibility for model specification. One of the two TMLE's for the parameter defined by the model correctly assumes that the binary outcome (e.g. disease or no disease) has a binomial distribution. This results in a double-robust (DR), efficient estimator of the parameter of interest in the model, but it is unstable, due to convergence problems with the log-binomial regression model, which is used for estimation. The second TMLE instead incorrectly assumes that the outcome is a count of events and follows a Poisson distribution. However, we apply the second TMLE to data in which the outcome is binary. In this case, the TMLE is no longer efficient, but it does achieve stability. It also remains DR - that is, the efficient score estimating function in the semi-parametric Poisson regression model is an unbiased DR estimating function for the parameter of interest in the semi-parametric conditional mean model, which does not assume a Poisson distribution. Consequently, this second TMLE is consistent and can provide correct inference. We refer to this latter TMLE as the "practical TMLE" of our parameter of interest when the outcome in truly binary, and we focus on this TMLE in our implementation and simulations in this paper. This chapter overlaps considerably with content found in Tuglus et al. [2011] but is presented differently.

- Chapter 4 provides theoretical details at a level much greater than in any previous chapters. Focusing on the third parameter of interest conditional relative risk in a semi-parametric multiplicative regression model, this chapter provides rigorous derivations of key features of the two corresponding TMLE's introduced in Chapter 3. In particular,

this chapter shows, for each of the TMLE's, how to construct the efficient score and efficient influence curve. It also presents the parametric fluctuation submodels for the TMLE step, including the so called clever covariates that define the submodels.

Chapter 2

Diagnosing and Responding to Violations in the Positivity Assumption

2.1 Introduction

Incomplete control of confounding is a well-recognized source of bias in causal effect estimation- measured covariates must be sufficient to control for confounding in order for causal effects to be identified based on observational data. The identifiability of causal effects further requires sufficient variability in treatment or exposure assignment within strata of confounders. The dangers of causal effect estimation in the absence of adequate data support have long been understood.Cochran [1957] More recent causal inference literature refers to the need for adequate exposure variability within confounder strata as the assumption of positivity or experimental treatment assignment. Robins [1986, 1987a, 1999] While perhaps less well-recognized than confounding bias, violations and near violations of the positivity assumption can increase both the variance and bias of causal effect estimates, and if undiagnosed can seriously threaten the validity of causal inferences.

Positivity violations can arise for two reasons. First, it may be theoretically impossible for individuals with certain covariate values to receive a given exposure of interest. For example, certain patient characteristics may constitute an absolute contraindication to receipt of a particular treatment. The threat to causal inference posed by such structural or theoretical violations of positivity does not improve with increasing sample size. Second, violations or near violations of positivity can arise in finite samples due to chance. This is a particular problem in small samples, but also occurs frequently in moderate to large samples when the treatment is continuous or can take multiple levels, or when the covariate adjustment set is large and/or contains continuous or multi-level covariates. Regardless of the cause, causal effects may be poorly or non-identified when certain subgroups in a finite sample do not receive some of the treatment levels of interest. In this paper, we will use the term "sparsity" to refer to positivity violations and near-violations arising from either of these causes, recognizing that other types of sparsity can also threaten valid inference.

In this chapter, we discuss the positivity assumption within a general framework for assessing the identifiability of causal effects. The causal model and target causal parameter are defined using a non-parametric structural equation model (NPSEM) and the positivity assumption is introduced as a key assumption needed for parameter identifiability. The counterfactual or potential outcome framework is then used to review estimation of the target parameter, assessment of the extent to which data sparsity threatens valid inference for this parameter, and practical approaches for responding to such threats. For clarity, we focus on a simple data structure in which treatment is assigned at a single time point. Concluding remarks generalize to more complex longitudinal data structures.

Data sparsity can increase both the bias and variance of a causal effect estimator; the extent to which each are impacted will depend on the estimator used. An estimator-specific diagnostic tool is thus needed to quantify the extent to which positivity violations threaten the validity of inference for a given causal effect parameter (for a given model, data-generating distribution, and finite sample). Wang et al. [2006a] proposed such a diagnostic based on the parametric bootstrap. Application of a candidate estimator to bootstrapped data sampled

from the estimated data generating distribution provides information about the estimator's behavior under a data generating distribution that is based on the observed data. The true parameter value in the bootstrap data is known and can be used to assess estimator bias. A large bias estimate can alert the analyst to the presence of a parameter that is poorly identified, an important warning in settings where data sparsity may not be reflected in the variance of the causal effect estimate.

Once bias due to violations in positivity have been diagnosed, the question remains how best to proceed with estimation. We review several approaches. Identifiability can be improved by extrapolating based on subgroups in which sufficient treatment variability does exist; however, such an approach requires additional parametric model assumptions. Alternative approaches for responding to sparsity include the following: restriction of the sample to those subjects for whom the positivity assumption is not violated (known as trimming); re-definition of the causal effect of interest as the effect of only those treatments that do not result in positivity violations (estimation of the effects of "realistic" or "intention to treat" dynamic regimes); restriction of the covariate adjustment set to exclude those covariates responsible for positivity violations; and, when the target parameter is defined using a marginal structural working model, use of a projection function that focuses estimation on areas of the data with greater support.

As we discuss, all of these approaches change the parameter being estimated by trading proximity to the original target of inference for improved identifiability. We advocate incorporation of this tradeoff into the effect estimator itself. This requires defining a family of parameters, the members of which vary in their proximity to the initial target and in their identifiability. An estimator can then be defined that selects among the members of this family according to some pre-specifed criteria.

2.1.1 Outline

The chapter is structured as follows. Section 2.2 introduces a non-parametric structural equation model for a simple point treatment data structure, defines the target causal parameter using a non-parametric marginal structural model, and discusses conditions for parameter identifiability with an emphasis on the positivity assumption. Section 2.3 reviews three classes of causal effect estimators and discusses the behavior of these estimators in the presence of positivity violations. Section 2.4 reviews approaches for assessing threats to inference arising from positivity violations, with a focus on the parametric bootstrap. Section 2.5 investigates the performance of the parametric bootstrap as a diagnostic tool using simulated data. Section 2.6 then applies the diagnostic tool to a real data example. Section 2.7 reviews methods for responding to positivity violations once they have been diagnosed, and integrates these methods into a general approach to sparsity that is based on defining a family of parameters. Section 2.8 offers some concluding remarks and advocates a systematic approach to possible violations in positivity.

2.2 Framework for Causal Effect Estimation

We proceed from the basic premise that model assumptions should honestly reflect investigator knowledge. The non-parametric structural equation model (NPSEM) framework of Pearl provides a systematic approach for translating background knowledge into a causal model and corresponding statistical model, defining a target causal parameter, and assessing the identifiability of that parameter. Pearl [2000] We illustrate this approach using a simple point treatment data structure. We minimize notation by focusing on discrete-valued random variables.

2.2.1 Model

Let W denote a set of baseline covariates on a subject, let A denote a treatment or exposure variable, and let Y denote an outcome. Specify the following structural equation model (with random input $U \sim P_U$):

$$\begin{aligned} W &= f_W(U_W) \\ A &= f_A(W, U_A) \\ Y &= f_Y(W, A, U_Y), \end{aligned} \tag{2.1}$$

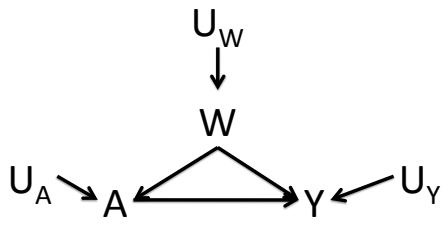
where $U = (U_W, U_A, U_Y)$ denotes the set of background factors that deterministically assign values to (W, A, Y) according to functions (f_W, f_A, f_Y). Each of the equations in this model is assumed to represent a mechanism that is autonomous, in the sense that changing or intervening on the equation will not affect the remaining equations, and that is functional, in the sense that the equation reflects assumptions about how the observed data were in fact generated by Nature. In addition, each of the equations is non-parametric, in the sense that its specification does not require assumptions regarding the true functional form of the underlying causal relationships. However, if aspects of the functional form of any of these equations are known based on background knowledge, such knowledge can be incorporated into the model.

A causal graph is derived from a non-parameteric structural equation model by connecting each observed variable to its "parents" (the subset of covariates found in the right hand side of the corresponding structural equation) with arrows emanating from the parents. The causal graph corresponding to Model (2.1) is given in Figure 2.1. The background factors U are assumed to be jointly independent in this particular model; or in other words, the model is assumed to be Markov. Pearl [2000] This assumption is encoded in the absence of double headed arrows between the elements of U in Figure 2.1. The NPSEM framework can also be applied to non-Markov models.

Let the observed data consist of n i.i.d. observations $O_1, \ldots, O_n$ of

$$O = (W, A, Y) \sim P_0.$$

Figure 2.1: Causal graph for non-parametric structural equation model



Causal model (2.1) places no restrictions on the allowed distributions for P_0, and thus implies a non-parametric statistical model.

2.2.2 Target Causal Parameter

A causal effect can be defined in terms of the joint distribution of the observed data under an intervention on one or more of the structural equations, or equivalently, under an intervention on the causal graph. For example, consider the post-intervention distribution of Y under an intervention on the structural model to set $A = a$. Such an intervention corresponds to replacing $A = f_A(W, U_A)$ with $A = a$ in the structural model (2.1), as follows:

$$\begin{aligned} W &= f_W(U_W) \\ A &= a \\ Y &= f_Y(W, a, U_Y). \end{aligned} \tag{2.2}$$

The counterfactual outcome that a given subject with background factors u would have had if he or she were to have received treatment level a is denoted $Y_a(u)$.Neyman [1923], Rubin [1974] This counterfactual can be derived as the solution to the structural equation f_Y in equation system (2.2) within input $U = u$.

Let F_X denote the distribution of $X = (W, (Y_a : a \in \mathcal{A}))$, where $\mathcal{A}$ denotes the possible values that the treatment variable can take (e.g. $\{0, 1\}$ for a binary treatment). F_X describes the joint distribution of the baseline covariates and counterfactual outcomes under a range of interventions on treatment variable A. A causal effect can be defined as some function of

F_X. For example, a common target parameter for binary A is the average treatment effect

$$E_{F_X}(Y_1 - Y_0), \tag{2.3}$$

or the difference in expected counterfactual outcome if every subject in the population had received versus had not received treatment.

Alternatively, an investigator may be interested in estimating the average treatment effect separately within certain strata of the population and/or for non-binary treatments. Specification of a marginal structural model (a model on the conditional expectation of the counterfactual outcome given effect modifiers of interest) provides one option for defining the target causal parameter in such cases.Robins [1999, 1998, 1999] Marginal structural models take the following form:

$$E_{F_X}(Y_a \mid V) = m(a, V \mid \beta), \tag{2.4}$$

where $V \subset W$ denotes the strata in which one wishes to estimate a conditional causal effect. For example, one might specify the following model:

$$m(a, V \mid \beta) = \beta_1 + \beta_2 a + \beta_3 V + \beta_4 aV.$$

For a binary treatment $\mathcal{A} \in \{0, 1\}$, such a model implies an average treatment effect within stratum $V = v$ equal to $\beta_2 + \beta_4 v$.

The true functional form of $E_{F_X}(Y_a \mid V)$ will generally not be known. One option is to assume that the parametric model $m(a, V \mid \beta)$ is correctly specified, or in other words that $E_{F_X}(Y_a \mid V) = m(a, V \mid \beta)$ for some value β. Such an approach, however, can place additional restrictions on the allowable distributions of the observed data and thus change the statistical model. In order to respect the premise that the statistical model should faithfully reflect the limits of investigator knowledge and not be altered in order to facilitate definition of the target parameter, we advocate an alternative approach in which the target causal parameter is defined using a non-parametric marginal structural model. Under this approach the target parameter β is defined as the projection of the true causal curve $E_{F_X}(Y_a \mid V)$ onto the specified model $m(a, V \mid \beta)$ according to some projection function $h(a, V)$:

$$\beta(F_X, m, h) = \underset{\beta}{\operatorname{argmin}}\, E_{F_X}\left[\sum_{a \in \mathcal{A}} (Y_a - m(a, V|\beta))^2 h(a, V)\right] .Neugebauerandvan\ derLaan\ [2007] \tag{2.5}$$

When $h(a, V) = 1$, the target parameter β corresponds to an unweighted projection of the entire causal curve onto the model $m(a, V|\beta)$; alternative choices of h correspond to placing greater emphasis on specific parts of the curve (i.e. on certain (a, V) values).

Use of a non-parametric marginal structural model such as (2.5) is attractive because it allows the target causal parameter to be defined within the original statistical model. However, this approach by no means absolves the investigator from careful consideration of marginal structural model specification. A poorly specified model $m(a, V|\beta)$ may result in a

target parameter that provides a poor summary of the features of the true causal relationship that are of interest.

In the following sections we discuss the parameter $\beta(F_X, m, 1)$ as the target of inference, corresponding to a focus on estimation of the treatment-specific mean for all levels $a \in \mathcal{A}$ within strata of V as projected onto model m, with projection $h(a, V) = 1$ chosen to reflect a focus on the entire causal curve. To simplify notation we use β to refer to this target parameter unless otherwise noted.

2.2.3 Identifiability

We assess whether the target parameter β of the counterfactual data distribution F_X is identified as a parameter of the observed data distribution P_0 under causal Model (2.1). Because Model (2.1) is Markov, we have that

$$P_{F_X}(Y_a = y) = \sum_w P_0(Y = y|W = w, A = a)P_0(W = w), \tag{2.6}$$

identifying the target parameter β according to projection (2.5).Pearl [2000] This identifiability result is often referred to as the G-computation formula.Robins [1986, 1987a,b] The weaker assumption of randomization, or the assumption that A and Y_a are conditionally independent given W, is also sufficient for identifiability result (2.6) to hold.

Randomization Assumption:

$$A \coprod Y_a|W \text{ for all } a \in \mathcal{A}. Robins \text{ [1986, 1987}a, b\text{]} \tag{2.7}$$

Whether or not a given structural model implies that assumption (2.7) holds can be assessed directly from the graph through the back door criterion. Pearl [2000]

The need for experimentation in treatment assignment

The G-computation formula (2.6) is only a valid formula if the conditional distributions in the formula are well-defined. Let $g_0(a \mid W) \equiv P_0(A = a \mid W), a \in \mathcal{A}$ denote the conditonal distribution of treatment variable A under the observed data distribution P_0. If one or more treatment levels of interest do not occur within some covariate strata, the conditional probability $P_0(Y = y|A = a, W = w)$ will not be well-defined for some value(s) (a, w) and the identifiability result (2.6) will break down.

A simple example provides intuition into the threat to parameter identifiability posed by sparsity of this nature. Consider an example in which $W = I(woman)$, A is a binary treatment, and no women are treated ($g_0(1|W = 1) = 0$). In this data generating distribution there is no information regarding outcomes among treated women. Thus, as long as there are women in the target population (i.e. $P_0(W = 1) > 0$), the average treatment effect $E_{F_X}(Y_1 - Y_0)$ will not be identified without additional parametric assumptions.

This simple example illustrates that a given causal parameter under a given model may be identified for some joint distributions of the observed data but not for others. An additional assumption is thus needed to ensure identfiability. We begin by presenting the strong version of this assumption, needed for the identification of $P_{F_X}((Y_a = y, W = w) : a, y, w)$ in a non-parametric model.

Strong Positivity Assumption:

$$\inf_{a \in \mathcal{A}} g_0(a \mid W) > 0, \text{- a.e.} \tag{2.8}$$

The strong positivity assumption, or assumption of experimental treatment assignment (ETA), states that each possible treatment level occurs with some positive probability within each strata of W.

Parametric model assumptions may allow the positivity assumption to be weakened. In the example above, an assumption that the treatment effect is the same among treated men and women would result in identification of the average treatment effect (2.3) based on extrapolation from the estimated treatment effect among men (assuming that other identifiability assumptions were met). Parametric model assumptions of this nature are particularly dangerous, however, because they extrapolate to regions of the joint distribution of (A, W) that are not supported by the data. Such assumptions should be approached with caution and adopted only when they have a solid foundation in background knowledge.

In addition to being model-specific, the form of the positivity assumption needed for identifiability is parameter-specific. Many target causal parameters require much weaker versions of positivity than (2.8). To take one simple example, if the target parameter is $E(Y_1)$, the identifiability result only requires that $g_0(1|W) > 0$ hold; it doesn't matter if there are some strata of the population in which no one was treated. Similarly, the identifiability of $\beta(F_X, m, h)$, defined using a marginal structural model, relies on a weaker positivity assumption.

Positivity Assumption for $\beta(F_X, h, m)$:

$$\sup_{a \in \mathcal{A}} \frac{h(a, V)}{g(a|W)} < \infty, \text{- a.e.} \tag{2.9}$$

Choice of projection function $h(a, V)$ used to define the target parameter thus has implications for how strong an assumption on positivity is needed for identifiability. In Section 2.7 we consider specification of alternative target parameters that allow for weaker positivity assumptions than (2.8), including parameters indexed by alternative choices of $h(a, V)$. For now we focus on the target parameter β indexed by the choice $h(a, V) = 1$ and note that (2.8) and (2.9) are equivalent for this parameter.

Once a target parameter has been specified, an assessment of its identifiability should precede estimation. Causal graphs provide a tool for assessment of identifiability assumption (2.7); however, an additional tool is needed to assess threats to identifiability arising from positivity violations or near violations. Section 2.4 reviews approaches for diagnosing

such threats, with a focus on the parametric bootstrap. Because the impact of positivity violations is estimator-specific, we first review several common estimators of β and discuss their behavior in the face of sparsity.

2.3 Estimator-specific Behavior in the Face of Positivity Violations

Let $\Psi(P_0)$ denote the target parameter of the observed data distribution, which under the assumptions of randomization (2.7) and positivity (2.9) equals the target causal parameter $\beta(F_X, m, h)$. Estimators of this parameter are denoted $\hat{\Psi}(P_n)$, where P_n is the empirical distribution of a sample of n i.i.d observations from P_0. We use $Q_{0W}(w) \equiv P_0(W = w)$, $Q_{0Y}(y|A, W) \equiv P_0(Y = y|A, W)$, and $\bar{Q}_0(A, W) \equiv E_0(Y|A, W)$. Recall that $g_0(a|W) \equiv P_0(A = a|W)$. We review three classes of estimators $\hat{\Psi}(P_n)$ of β that employ estimators of distinct parts of the observed data likelihood. Maximum likelihood-based substitution estimators (also referred to as "G-computation" estimators) employ estimators of $Q_0 \equiv (Q_{0W}, \bar{Q}_0)$. Inverse probability weighted estimators employ estimators of g_0. Double robust estimators employ estimators of both g_0 and Q_0. A summary of these estimators is provided in Table 2.1. Their behavior in the face of positivity violations is illustrated in Section 2.5 and previous work. Neugebauer and van der Laan [2007, 2005], Bembom and van der Laan [2007], Moore et al. [2009], Cole and Hernan [2008]

We focus our discussion on bias in the point estimate of the target parameter β. While estimates of the variance of β can also be biased when data are sparse, methods exist to improve variance estimation. The non-parametric bootstrap provides one straightforward approach to variance estimation in setting where the central limit theorem may not apply as a result of sparsity; alternative approaches to correct for biased variance estimates are also possible.Rosenblum and van der Laan [2001] These methods will not, however, protect against misleading inference if the point estimate itself is biased.

2.3.1 The G-computation Estimator

The G-computation estimator $\hat{\Psi}_{Gcomp}(P_n)$ provides a mapping from the empirical data distribution P_n to a parameter estimate $\hat{\beta}_{Gcomp}$. $\hat{\Psi}_{Gcomp}(P_n)$ is a substitution estimator based on identifiability result (2.6). It is implemented based on an estimator of $Q_0 \equiv (Q_{0W}, \bar{Q}_0)$ and its consistency relies on the consistency of this estimator.Robins [1986, 1987a] Q_{0W} can generally be estimated based on the empirical distribution of W. However, even when positivity is not violated, the dimension of A, W is frequently too large for $\bar{Q}_0$ to be estimated simply by evaluating the mean of Y within strata of (A, W). Due to the curse of dimensionality, estimation of $\bar{Q}_0$ under a non-parametric or semi-parametric statistical model thus frequently requires data-adaptive approaches such as cross-validated loss-based learning.van der Laan and Dudoit [2003b], van der Laan et al. [2007], Hastie et al. [2009]

Table 2.1: Overview of three classes of causal effect estimator.

G-computation Estimator (Section 2.3.1)	
Needed for Implementation:	Estimator Q_n of Q_0
Needed for Consistency:	Q_n is a consistent estimator of Q_0
Response to Sparsity:	Extrapolates based on Q_n Sparsity can amplify bias due to model misspecification
IPTW Estimator (Section 2.3.2.)	
Needed for Implementation:	Estimator g_n of g_0
Needed for Consistency:	g_n is a consistent estimator of g_0 g_0 satisfies positivity
Response to Sparsity:	Does not extrapolate based on Q_n Sensitive to positivity violations and near violations
DR Estimators (Section 2.3.3.)	
Needed for Implementation:	Estimator g_n of g_0 and Q_n of Q_0
Needed for Consistency:	g_n is consistent or Q_n is consistent g_n converges to a distribution that satisfies positivity
Response to Sparsity:	Can extrapolate based on Q_n Without positivity, relies on consistency of Q_n

Given an estimator $\bar{Q}_n$ of $\bar{Q}_0$, the G-computation estimator can be implemented by generating a predicted counterfactual outcome for each subject under each possible treatment: $\hat{Y}_{a,i} = \bar{Q}_n(a, W_i)$ for $a \in \mathcal{A}$, $i = 1, ..., n$. The estimate $\hat{\beta}_{Gcomp}$ is then obtained by regressing $\hat{Y}_a$ on a and V according to the model $m(a, V \mid \beta)$, with weights based on the projection function $h(a, V)$.

When all treatment levels of interest are not represented within all covariate strata (i.e. assumption (2.8) is violated), some of the conditional probabilities in the non-parametric G-computation formula (2.6) will not be defined. A given estimate $\bar{Q}_n$ may allow the G-computation estimator to extrapolate based on covariate strata in which sufficient experimentation in treatment level does exist. Importantly, however, this extrapolation depends heavily on the model for $\bar{Q}_0$ and the resulting effect estimates will be biased if the model used to estimate $\bar{Q}_0$ is misspecified.

Moore et. al. illustrate the bias that can arise in the G-computation estimator when simple model fitting algorithms such as forward and backward selection are used to estimate $\bar{Q}_0(A, W)$. Moore et al. [2009] While more sophisticated model fitting techniques can improve estimator performance, they do not resolve the potential for data sparsity to result in substatial bias. One possible source of positivity violations is collinearity between a confounder or set of confounders and the treatment or exposure of interest. If data-adaptive methods are used to fit $\bar{Q}(A, W)$, covariates that are collinear or highly correlated with treatment may be dropped from a model in which treatment is forced. If these covariates are also confounders, resulting effect estimates will be biased.

Traditional Multivariable Approaches. A traditional approach to effect estimation

in many fields is to estimate $\bar{Q}_0 \equiv E_0(Y|A,W)$ using a multivariable regression model and to report the estimated coefficient on A (or some transformation of this coefficient, such as its exponentiated value) as the estimated causal effect. In some cases such an estimate is equivalent to the G-computation estimate. For example, if the target of inference is the average treatment effect for binary A, a traditional analysis might fit the model $\hat{E}(Y|A,W) = \hat{\beta}_0 + \hat{\beta}_1 A + k(W)$ and report an effect estimate of $\hat{\beta}_1$. In this case, $\hat{\beta}_1$ will be equivalent to $\hat{\beta}_{Gcomp}$ (assuming the same model is used for $\bar{Q}_n$ when implementing the G-computation estimator).

In many cases, however, the coefficient on A in the multivariable regression model used to estimate $\bar{Q}_0$ represents a distinct estimand. For example, for binary Y a common approach is to fit a logistic regression model such as $\hat{E}(Y|A,W) = 1/(1+\exp^{-(\hat{\beta}_0+\hat{\beta}_1 A+k(W))})$. Here $\exp(\hat{\beta}_1)$, which is commonly reported as the causal effect estimate of interest, is an estimate of the conditional odds ratio and is not equivalent to either the average treatment effect or the marginal odds ratio. If G-computation is used to estimate either of the latter two quantities then clearly the resulting estimates will not be equivalent. Traditional regression approaches can consistently estimate causal parameters when identifiability conditions are met and $\bar{Q}_n$ is correctly specified; however, care must be taken to ensure that the parameter estimated corresponds to the causal question of interest.

2.3.2 The Inverse Probability of Treatment Weighted Estimator

The IPTW estimator $\hat{\Psi}_{IPTW}(P_n)$ provides a mapping from the empirical data distribution P_n to a parameter estimate $\hat{\beta}_{IPTW}$ based on an estimator g_n of $g_0(A|W)$.Robins [1999], Robins et al. [2000] The estimator is defined as the solution in β to the following estimating equation:

$$0 = \sum_{i=1}^{n} \frac{h(A_i, V_i)}{g_n(A_i \mid W_i)} \frac{d}{d\beta}(m(A_i, V_i|\beta))(Y - m(A_i, V_i \mid \beta)), \tag{2.10}$$

where $h(A,V)$ is the projection function used to define the target causal parameter $\beta(F_X, m, h)$ according to (2.5). The IPTW estimator of β can be implemented as the solution to a weighted regression of the outcome Y on treatment A and effect modifiers V according to model $m(A,V|\beta)$, with weights equal to $\frac{h(A,V)}{g_n(A|W)}$. Consistency of $\hat{\Psi}_{IPTW}(P_n)$ requires that g_0 satisfies positivity and that g_n is a consistent estimator of g_0. As with $\bar{Q}_0$, g_0 can be estimated using loss-based learning and cross validation. Depending on choice of projection function, implementation may further require estimation of $h(A,V)$; however, the consistency of the IPTW estimator does not depend on consistent estimation of $h(A,V)$.

The IPTW estimator is particularly sensitive to bias due to data sparsity. Bias can arise due to structural positivity violations (positivity may not hold for g_0) or may occur because by chance certain covariate and treatment combinations are not represented or sparsely represented in a given finite sample. In the latter case, $g_n(a|W=w)$ will have values of zero or close to zero for some (a,w) even when positivity holds for g_0 and g_n is consistent.Wang

et al. [2006a], Neugebauer and van der Laan [2005], Bembom and van der Laan [2007], Cole and Hernan [2008], Moore et al. [2009] As fewer individuals within a given covariate stratum receive a given treatment, the weights of those rare individuals who do receive the treatment become more extreme. The disproportionate reliance of the causal effect estimate on the experience of a few unusual individuals can result in substantial finite sample bias.

While values of $g_n(a \mid W)$ remain positive for all $a \in \mathcal{A}$, elevated weights inflate the variance of the effect estimate and can serve as a warning that the data may poorly support the target parameter. However, as the number of individuals within a covariate stratum who receive a given treatment level shifts from few (each of whom receive a large weight and thus elevate the variance) to none, estimator variance can decrease while bias increases rapidly. In other words, when $g_n(a|W = w) = 0$ for some (a, w), the weight for a subject with $A = a$ and $W = w$ is infinity; however, as no such individuals exist in the dataset, the corresponding threat to valid inference will not be reflected in either the weights or in estimator variance.

Weight truncation. Weights are commonly truncated or bounded in order to improve the performance of the IPTW estimator in face of data sparsity.Wang et al. [2006a], Moore et al. [2009], Cole and Hernan [2008], Kish [1992], Bembom and van der Laan [2008] Weights are truncated at either a fixed or relative level (for example, at the 1st and 99th percentiles), thereby reducing the variance arising from large weights and limiting the impact of a few possibly non-representative individuals on the effect estimate. This advantage comes at a cost, however, in the form of increased bias due to misspecification of the treatment model g_n, a bias that does not decrease with increasing sample size. In Section 2.5, we use simulated data to illustrate the performance of the IPTW estimator under a range of values for weight truncation, illustrate how even in the face of sparsity, weight truncation can increase rather than decrease estimator mean squared error, and discuss how the parametric bootstrap can be used to approach truncation.

Stabilized Weights. Use of projection function $h(a, V) = 1$ implies the use of unstabilized weights. In contrast, stabilized weights, corresponding to a choice of $h(a, V) = g(a|V)$ (where $g(a|V)$ denotes $P_0(A = a|V)$) are generally recommended for the implementation of marginal structural model-based effect estimation. The choice of $h(a, V) = g(a|V)$ results in a weaker positivity assumption, according to (2.9), by allowing the IPTW estimator to extrapolate to sparse areas of the joint distribution of (A, V) using the model $m(a, V|\beta)$. For example, if A is an ordinal variable with multiple levels, $V = \{\}$, and the target parameter is defined using the model $m(a, V|\beta) = \beta_0 + \beta_1 a$, the IPTW estimator with stabilized weights will extrapolate to levels of A that are sparsely represented in the data by assuming a linear relationship between Y_a and $a \in \mathcal{A}$. We note, however, that when the target parameter β is defined using a non-parametric marginal structural model according to (2.5) (an approach that acknowledges that the model $m(A, V|\beta)$ may be misspecified), the use of stabilized versus unstabilized weights corresponds to a shift in the target parameter via choice of an alternative projection function.Neugebauer and van der Laan [2007]

2.3.3 Double Robust Estimators

Double robust (DR) approaches to estimation of β include the augmented inverse probability weighted estimator (A-IPTW) and targeted maximum likelihood estimator (TMLE) (which for the target parameter $\beta(F_X, h, m)$ corresponds to the extended double robust parametric regression estimator of Sharfstein *et. al.*).Robins [1999], Robins and Rotnitzky [2001], Robins [2002], Scharfstein et al. [1999], van der Laan and Rubin [2006], Rosenblum and van der Laan [2010a] Implementation of the double robust estimators requires estimators of both Q_0 and g_0; as with the IPTW and G-computation estimators, a non-parametric loss-based approach can be employed in the estimation of both. An implementation of the TMLE estimator of the average treatment effect $E(Y_1 - Y_0)$ for binary A is available in the R package `tmleLite`; an implementation of the A-IPTW estimator in the point treatment setting is available in the R package `cvDSA` (both available at http://www.stat.berkeley.edu/ laan/Software/index.html). Prior literature provides further details regarding implementation and theoretical properties.Robins [1999], Neugebauer and van der Laan [2007, 2005], Robins and Rotnitzky [2001], Scharfstein et al. [1999], van der Laan and Rubin [2006], Rosenblum and van der Laan [2010a]

Double robust estimators remain consistent if either 1) g_n is a consistent estimator of g_0 and g_0 satisfies positivity; or, 2) Q_n is a consistent estimator of Q_0 and g_n converges to a distribution g^* that satisfies positivity. Thus when positivity holds, these estimators are truly double robust, in the sense that consistent estimation of either g_0 or Q_0 results in a consistent estimator. When positivity fails, however, the consistency of the double robust estimators relies entirely on consistent estimation of Q_0. In the setting of positivity violations, double robust estimators are thus faced with the same vulnerabilities as the G-computation estimator.

In addition to illustrating how positivity violations increase the vulnerability of double robust estimators to bias resulting from inconsistent estimation of Q_0, these asymptotic results have practical implications for the implementation of the double robust estimators. Specifically, they suggest that use of an estimator g_n that satisfies positivity (or in other words, that yields predicted values in $[0 + \gamma, 1 - \gamma]$ where γ is some small number) can improve finite sample performance. One way to achieve such bounds is by truncating the predicted probabilities generated by g_n, similar to the process of weight truncation described for the IPTW estimator.

Alternative double robust estimators are available that make more sophisticated choices in estimating g_0. In particular, the collaborative targeted maximum likelihood estimator (C-TMLE) selects an estimator g_n aimed at optimizing estimation of the target parameter as assessed by the targeted log likelihood. In particular this implies that the C-TMLE estimator includes in the fit of g_n only those covariates that improve estimation of the target.van der Laan and Gruber [2009a] However, when the target parameter is poorly identified due to positivity violations, C-TMLE may be forced to accept significant bias in its aim to optimize mean squared error for the target parameter. Diagnostic procedures remain essential to alert

the analyst that such a tradeoff is occurring.

2.4 Diagnosing Bias Due to Positivity Violations

Positivity violations can result in substantial bias, with or without a corresponding increase in variance, regardless of the causal effect estimator used. Practical methods are thus needed to diagnose and quantify estimator-specific positivity bias for a given model, parameter and sample. Cole and Hernan suggest a range of informal diagnostic approaches when the IPTW estimator is applied.Cole and Hernan [2008] Basic descriptive analyses of treatment variability within covariate strata can be helpful; however, this approach quickly becomes unwieldy when the covariate set is moderately large and includes continuous or multi-level variables. Examination of the distribution of the estimated weights can also provide useful information as near violations of the positivity assumption will be reflected in large weights. As noted by these authors and discussed above, however, well-behaved weights are not sufficient in themselves to ensure the absence of positivity violations.

An alternative formulation is to examine the distribution of the estimated propensity score values given by $g_n(a|W)$ for $a \in \mathcal{A}$. Values of $g_n(a|W)$ close to 0 for any a constitute a warning regarding the presence of positivity violations. We note that examination of the propensity score distribution is a general approach not restricted to the IPTW estimator. However, while useful in diagnosing the presence of positivity violations, examination of the estimated propensity scores does not provide any quantitative estimate of the degree to which such violations are resulting in estimator bias and may pose a threat to inference. The parametric bootstrap can be used to provide an optimistic bias estimate specifically targeted at bias caused by positivity violations and near-violations. Wang et al. [2006a]

2.4.1 The Parametric Bootstrap as a Diagnostic Tool

We focus on the bias of estimators that target a parameter of the observed data distribution; this target observed data parameter is equal under the randomization assumption (2.7) to the target causal parameter. (Divergence between the target observed data parameter and target causal parameter when (2.7) fails is a distinct issue not addressed by the proposed diagnostic.) The bias in an estimator is the difference between the true value of the target parameter of the observed data distribution and the expectation of the estimator applied to a finite sample from that distribution:

$$Bias(\hat{\Psi}, P_0, n) = E_{P_0}\hat{\Psi}(P_n) - \Psi(P_0),$$

where we recall that $\Psi(P_0)$ is the target observed data parameter, $\hat{\Psi}(P_n)$ is an estimator of that parameter (which may be a function of g_n, Q_n or both), and P_n denotes the empirical distribution of a sample of n i.i.d observations from the true observed data distribution P_0.

Bias in an estimator can arise due to a range of causes. First, the estimators g_n and/or Q_n may be inconsistent. Second, g_0 may not satisfy the positivity assumption. Third, consistent estimators g_n and/or Q_n may still have substantial finite sample bias. This latter type of finite sample bias arises in particular due to the curse of dimensionality in a non-parametric or semi-parametric model when g_n and/or Q_n are data-adaptive estimators, although it can also be substantial for parametric estimators. Fourth, estimated values of g_n may be equal or close to zero or one, despite use of a consistent estimator g_n and a distribution g_0 that satisfies positivity. The relative contribution of each of these sources of bias will depend on the model, the true data generating distribution, the causal effect estimator, and the finite sample.

The parametric bootstrap provides a tool that allows the analyst to explore the extent to which bias due to any of these causes is affecting a given parameter estimate. The parametric bootstrap-based bias estimate is defined as:

$$\widehat{Bias}_{PB}(\hat{\Psi}, \hat{P}_0, n) = E_{\hat{P}_0}\hat{\Psi}(P_n^{\#}) - \Psi(\hat{P}_0), \tag{2.11}$$

where $\hat{P}_0$ is an estimate of P_0 and $P_n^{\#}$ is the empirical distribution of a bootstrap sample obtained by sampling from $\hat{P}_0$. In other words, the parametric bootstrap is used to sample from an estimate of the true data generating distribution, resulting in multiple simulated data sets. The true data generating distribution and target parameter value in the bootstrapped data are known. A candidate estimator is then applied to each bootstrapped data set and the mean of the resulting estimates compared with the known "truth" (i.e. the true parameter value for the bootstrap data generating distribution).

We focus on a particular algorithm for parametric bootstrap-based bias estimation, which specifically targets the component of estimator-specific finite sample bias due to violations and near violations of the positivity assumption. The goal is not to provide an accurate estimate of bias, but rather to provide a diagnostic tool that can serve as a "red flag" warning that positivity bias may pose a threat to inference. The distinguishing characteristic of the diagnostic algorithm is its use of an estimated data generating distribution $\hat{P}_0$ that both approximates the true P_0 as closely as possible and is compatible with the estimators $\bar{Q}_n$ and/or g_n used in $\hat{\Psi}(P_n)$. In other words, $\hat{P}_0$ is chosen such that the estimator $\hat{\Psi}$ applied to bootstrap samples from $\hat{P}_0$ is guaranteed to be consistent unless g_0 fails to satisfy the positivity assumption or g_n is truncated. As a result, the parametric bootstrap provides an optimistic estimate of finite sample bias, in which bias due to model misspecification other than truncation is eliminated.

We refer informally to the resulting bias estimate as $ETA.Bias$ because in many settings it will be predominantly composed of bias from the following sources: 1) violation of the positivity assumption by g_0; 2) truncation, if any, of g_n in response to positivity violations; and, 3) finite sample bias arising from values of g_n close to zero or one (sometime referred to as practical violations of the positivity assumption). The term $ETA.Bias$ is imprecise because the bias estimated by the proposed algorithm will also capture some of the bias

in $\hat{\Psi}(P_n)$ due to finite sample bias of the estimators g_n and $\bar{Q}_n$ (a form of sparsity only partially related to positivity). Due to the curse of dimensionality, the contribution of this latter source of bias may be substantial when g_n and/or Q_n are data-adaptive estimators in a non-parametric or semi-parametric model. However, the proposed diagnostic algorithm will only capture a portion of this bias because, unlike P_0, $\hat{P}_0$ is guaranteed to have a functional form that can be well-approximated by the data-adaptive algorithms employed by g_n and Q_n.

The diagnostic algorithm for $ETA.Bias$ is implemented as follows.

Step 1. Estimate P_0. Estimation of P_0 requires estimation of Q_{0W}, g_0, and Q_{0Y}, (i.e. estimation of $P_0(W = w)$, $P_0(A = a|W = w)$, and $P_0(Y = y|A = a, W = w)$ for all (w, a, y)). We define $Q_{\hat{P}_0 W} = Q_{P_n W}$ (or in other words, use an estimate based on the empirical distribution of the data), $g_{\hat{P}_0} = g_n$, and $\bar{Q}_{\hat{P}_0} = \bar{Q}_n$. Note that the estimators $Q_{P_n W}$, g_n, and $\bar{Q}_n$ were all needed for implementation of the IPTW, G-compuation, and DR estimators; the same estimators can be used here. Additional steps may be required to estimate the entire conditional distribution of Y given (A, W) (beyond the estimate of its mean given by $\bar{Q}_n$). The true target parameter for the known distribution $\hat{P}_0$ is only a function of $Q_n = (Q_{P_n W}, \bar{Q}_n)$, and $\Psi(\hat{P}_0)$ is the same as the G-computation estimator (using Q_n) applied to the observed data:

$$\Psi(\hat{P}_0) = \hat{\Psi}_{Gcomp}(P_n).$$

Step 2. Generate $P_n^{\#}$ by sampling from $\hat{P}_0$. In the second step, we assume that $\hat{P}_0$ is the true data generating distribution. Bootstrap samples $P_n^{\#}$, each with n i.i.d observations, are generated by sampling from $\hat{P}_0$. For example, W can be sampled from the empirical, a binary A can be generated as a Bernoulli with probability $g_n(1|W)$, and a continuous Y can be generated by adding a $N(0, 1)$ error to $\bar{Q}_n(A, W)$ (alternative approaches are also possible).

Step 3. Estimate $E_{\hat{P}_0}\hat{\Psi}(P_n^{\#})$. Finally, the estimator $\hat{\Psi}$ is applied to each bootstrap sample. Depending on the estimator being evaluated, this step involves applying the estimators g_n, Q_n or both to each bootstrap sample. If Q_n and/or g_n are data-adaptive estimators, the corresponding data-adaptive algorithm should be rerun in each bootstrap sample; otherwise, the coefficients of the corresponding models should be refit. $ETA.Bias$ is calculated by comparing the mean of the estimator $\hat{\Psi}$ across bootstrap samples ($E_{\hat{P}_0}\hat{\Psi}_{IPTW}(P_n^{\#})$) with the true value of the target parameter under the bootstrap data generating distribution ($\Psi(\hat{P}_0)$).

The parametric bootstrap-based diagnostic applied to the IPTW estimator is available as an R function `check.ETA` in the `cvDSA` package.Wang et al. [2006a] The routine takes the original data as input and performs bootstrap simulations under user-specified information such as functional forms for $m(a, V \mid \beta)$, g_n and Q_n. Application of the bootstrap to the IPTW estimator offers one particularly sensitive assessment of positivity bias because, unlike the G-computation and double robust estimators, the IPTW estimator can not extrapolate based on $\bar{Q}_n$. However, this approach can be applied to any causal effect estimator, including

estimators introduced in Section 2.7 that trade off identifiability for proximity to the target parameter. In assessing the threat posed by positivity violations the bootstrap should ideally be applied to both the IPTW estimator and the estimator of choice.

Remarks on interpretation of the bias estimate. We caution against using the parametric bootstrap for any form of bias correction. The true bias of the estimator is $E_{P_0}\hat{\Psi}(P_n) - \Psi(P_0)$, while the parametric bootstrap estimates $E_{\hat{P}_0}\hat{\Psi}(P_n^{\#}) - \Psi(\hat{P}_0)$. The performance of the diagnostic thus depends on the extent to which $\hat{P}_0$ approximates the true data generating distribution. This suggests the importance of using flexible data-adaptive algorithms to estimate P_0. Regardless of estimation approach, however, when the target parameter $\Psi(P_0)$ is poorly identified due to positivity violations $\Psi(\hat{P}_0)$ may be a poor estimate of $\Psi(P_0)$. In such cases one would not expect the parametric bootstrap to provide a good estimate of the true bias. Further, the *ETA.Bias* implementation of the parametric bootstrap provides a deliberately optimistic bias estimate by excluding bias due to model misspecifcation for the estimators g_n and $\bar{Q}_n$.

Rather, the parametric bootstrap is proposed as a diagnostic tool. Even when the data generating distribution is not estimated consistently, the bias estimate provided by the parametric bootstrap remains interpretable in the world where the estimated data generating mechanism represents the truth. If the estimated bias is large, an analyst who disregards the implied caution is relying on an unsubstantiated hope that first, he or she has inconsistently estimated the data generating distribution but still done a reasonable job estimating the causal effect of interest; and second, the true data generating distribution is less affected by positivity (and other finite sample) bias than is the analyst's best estimate of it.

The threshold level of *ETA.Bias* that is considered problematic will vary depending on the scientific question and the point and variance estimates of the causal effect. With that caveat, we suggest the following two general situations in which *ETA.Bias* can be considered a "red flag" warning: 1) when *ETA.Bias* is of the same magnitude as (or larger than) the estimated standard error of the estimator; and, 2) when the interpretation of a bias-corrected confidence interval would differ meaningfully from initial conclusions.

Use of a data-adaptive algorithm for $\bar{Q}_n$ may result in exclusion of those elements of W responsible for positivity violations. Bootstrap data sampled from the resulting estimate $\hat{P}_0$ will contain less sparsity than is present in the true data generating distribution, resulting in an underestimate of bias due to positivity violations. One approach to improving the sensitivity of the diagnostic in such settings is to force the estimator $\bar{Q}_n(A, W)$ to include all W known or thought to contribute to positivity violations. The estimated propensity score provides a convenient dimension reduction of exactly those W. Thus a more comprehensive approach to identifying threats to inference due to positivity bias could involve implementing the bootstrap-based *ETA.Bias* diagnostic using several estimators Q_n, including an estimator that forces inclusion of A but allows W to be selected data adaptively and an estimator that forces inclusion of both A and the propensity score but allows W to be selected data-adaptively. Finally, when the targeted maximum likelihood estimator

is implemented, the bootstrap can sample from the targeted estimate of the likelihood it provides, an estimate in which Q_n is already a function of the propensity score. We demonstrate the propensity score-based approaches in Section 2.5; however, the performance of the diagnostic when data-adaptive approaches are used and positivity violations are present, as well as the relative performance of various approaches to improving diagnostic performance in such settings, should be investigated further.

2.5 Simulations

Data were simulated under three data generating distributions with different degrees and sources of positivity violations. In each set of simulations, four estimators described in Section 2.3, G- computation, IPTW, A-IPTW, and TMLE, were applied. (Specifically, TMLE was implemented with a logistic fluctuation for continuous and binary Y.)Gruber and van der Laan [2010a] For each simulation, each estimator was implemented using a range of approaches to estimate g_0 and Q_0. Both the behavior of the estimator and the performance of the parametric bootstrap as a diagnostic tool were investigated under each scenario. The objectives of these simulations were (1) to demonstrate how different estimators are affected differently by violations of the positivity assumption; (2) to demonstrate the value and limitations of the bootstrap-based diagnostic in different settings; and (3) to illustrate how the diagnostic might be used in practice to inform interpretation of results. We provide selected simulation results here; additional results together with simulation code are available at http://www.stat.berkeley.edu/ laan/Software/index.html.

2.5.1 Data Generating Distributions

All three simulations used a binary A, and targeted the same causal parameter, $E(Y_1-Y_0)$ or the average treatment effect. This target parameter is a special case of $\beta(F_X, m, h)$ corresponding to $V = \{\}$ and use of marginal structural model $m(a|\beta) = \beta_0 + \beta_1 a$, and a case in which G-computation corresponds to traditional regression-based adjustment. The true target parameter value $\Psi(P_0) = \beta_1$.

The simulations were based, to varying degrees, on a data generating distribution used by Freedman and Berk.Freedman and Berk [2008] Two baseline covariates, $W = (W_1, W_2)$, were generated bivariate normal, $N(\mu, \Sigma)$, with $\mu_1 = 0.5$, $\mu_2 = 1$, and $\Sigma = \begin{bmatrix} 2 & 1 \\ 1 & 1 \end{bmatrix}$. The true conditional expectation of Y, given A and W, $\bar{Q}_0(A, W) \equiv E_0(Y|A, W)$ is given by:

$$\bar{Q}_0(A, W) = 1 + A + W_1 + 2W_2,$$

and Y was generated as $\bar{Q}_0(A, W) + U$, with $U \sim N(0, 1)$. The true value of the target parameter $\Psi(P_0) = 1$. The true treatment mechanism, $g_0(1|W) \equiv P_0(A = 1|W)$ is given by:

$$g_0(1|W) = \Phi(0.5 + 0.25W_1 + 0.75W_2),$$

where Φ is the CDF of the standard normal distribution. In other words, the treatment mechanism, or conditional probability of treatment given covariates, was based on a probit model.

Simulation 1: For our first simulation, we modified g_0 to reduce the extent of positivity violations by multiplying all coefficients in g_0 by 0.3. Therefore, the true treatment mechanism in Simulation 1 is given by:

$$g_0(1|W) = \Phi(0.3(0.5 + 0.25W_1 + 0.75W_2)).$$

With this treatment mechanism, $g_0 \in [0.48, 0.92]$. We generated 250 samples of size 1000 for this simulation.

Simulation 2: Simulation 2 is identical to Freedman and Berk's original simulation described above. Again we generated 250 samples of size 1000. In this simulation, $g_0 \in [0.001, 1]$.

Simulation 3: For this simulation, $W1 \sim N(0.5, 1)$ and $W2 \sim Bernoulli(0.5)$. We varied $\bar{Q}_0(A, W)$ such that:

$$\bar{Q}_0(1, W) = expit(-1 + 5A + W_1 + 10W_2).$$

Binary Y was generated as a Bernoulli trial with probability $\bar{Q}_0(1, W)$. The target parameter $E(Y_1 - Y_0)$ for binary Y corresponds to the risk difference. For this simulation, $\Psi(P_0) = 0.29$. The treatment mechanism for this simulation is given by:

$$g_0(1|W) = expit(-3 - 1W_1 + 9W_2).$$

Binary A was generated as a Bernoulli trial with probability $g_0(1|W)$. Under this treatment mechanism A and $W2$ are collinear, with correlation 0.95 and $g_0 \in [0.001, 1]$. For this simulation, we generated 250 samples of size 200 instead of size 1000. The smaller sample size increased the sparsity in the data.

2.5.2 Investigation of Estimator Behavior and the Performance of the Parametric Bootstrap-based Diagnostic

The bias, variance, and mean squared error of each estimator were estimated by applying the estimator to 250 samples drawn from the three data generating distributions above. For Simulations 1 and 2, each of the four estimators was implemented with each of the following three approaches: 1) use of a correctly specified model to estimate both $\bar{Q}_0$ and g_0 (a specification referred to as "$Qcgc$"); 2) use of a correctly specified model to estimate $\bar{Q}_0$ but

omission of W_2 from the model used to estimate g_0 ("$Qcgm$"); and, 3) omission of W_2 from $\bar{Q}_n$ while correctly specifying the model used to estimate g_0 ("$Qmgc$"). In Simulation 3, each of the four estimators was implemented using correctly specified models for both g_0 and $\bar{Q}_0$ ($Qcgc$), and using forward stepwise selection based on AIC to estimate both $\bar{Q}_0$ and g_0, using the R function `step` and forcing A to be included in $\bar{Q}_n$ ("$Qdgd1$") . The double robust and IPTW estimators were further implemented using the following sets of bounds for the values of g_n: $[0, 1]$ (or no bounding), $[0.025, 0.975]$, $[0.05, 0.95]$, and $[0.1, 0.9]$. For the IPTW estimator, the latter three bounds correspond to truncation of the unstabilized weights at $[1.03, 40]$, $[1.05, 20]$, and $[1.11, 11.1]$.

The parametric bootstrap was then applied to estimate $ETA.Bias$ for 10 of the 250 samples from each of the three simulations. For each sample and for each model specification ($Qcgc$, $Qmgc$ and $Qcgm$ for Simulations 1 and 2; and $Qcgc$ and $Qdgd1$ for Simulation 3), the estimates Q_n and g_n were used to draw 1000 parametric bootstrap samples. Specifically, W was drawn from the empirical distribution for that sample; A was generated as a series of Bernoulli trials with probability $g_n(1|W)$, and Y was generated either by adding a $N(0, 1)$ error to $\bar{Q}_n(A, W)$ (for continuous Y in Simulations 1 and 2) or as a series of Bernoulli trials with probability $\bar{Q}_n(1|A, W)$ (for binary Y in Simulation 3). Each candidate estimator was then applied to each bootstrap sample. In Simulation 3, an alternative implementation of the diagnostic based on including the propensity score in $\bar{Q}_n$ was also applied ("$Qdgd2$"). Specifically, the stepwise algorithm was forced to retain both A and the estimated propensity score $g_n(1|W)$ as covariates in the estimate $\bar{Q}_n$ used to generate the bootstrap samples.

For the specifications $Qcgc$, $Qmgc$ and $Qcgm$, the models used to estimate g_0 and $\bar{Q}_0$ were held fixed across bootstrap samples and their coefficients refit in each bootstrap sample. For the data-adaptive approaches $Qdgd1$ and $Qdgd2$, the stepwise selection algorithm was rerun in each bootstrap sample, and was forced to retain A in $\bar{Q}_n$. $ETA.Bias$ was estimated for each of the 10 samples as the difference between the mean of the bootstrapped estimator and the initial G-computation estimate $\Psi(\hat{P}_0) = \hat{\Psi}_{Gcomp}(P_n)$ in that sample.

2.5.3 Results: Simulation 1

In this simulation the positivity assumption is not violated, and as expected, all four estimators performed well when correctly specified models were used to estimate g_0 and $\bar{Q}_0$. The bias, variance, and MSE for each estimator are shown in Table 2.2. As described in Section 2.3, misspecification of the model used to estimate $\bar{Q}_0$ introduced bias in the G-computation estimator, misspecification of the model used to estimate g_0 introduced bias in the IPTW estimator, and the double robust estimators remained minimally biased if the model for either $\bar{Q}_0$ or g_0 was correctly specified.

Table 2.3 reports the mean and variance of the estimated $ETA.Bias$ for each estimator and model specification across 10 of the 250 original samples. Consistent with the results in Table 2.2, the estimated $ETA.Bias$ was minimal and varied little across the 10 samples. The parametric bootstrap would not have raised a red flag for any of the estimators in this

Table 2.2: Performance of estimators by specification in Simulation 1: g_0 in [0.48,0.92], shown for unbounded g_n only. Results are based on 250 samples of size 1000.

	Qcgc			**Qcgm**			**Qmgc**		
	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE
G-COMP	1.5e-03	5.9e-03	5.9e-03	1.5e-03	5.9e-03	5.9e-03	2.6e-01	1.9e-02	8.5e-02
IPTW	6.0e-03	9.2e-03	9.2e-03	2.6e-01	2.1e-02	9.0e-02	6.0e-03	9.2e-03	9.2e-03
A-IPTW	2.6e-04	6.2e-03	6.2e-03	5.9e-04	6.0e-03	6.0e-03	7.2e-04	6.7e-03	6.7e-03
TMLE	-6.7e-06	6.2e-03	6.2e-03	3.9e-04	6.0e-03	6.0e-03	5.0e-04	6.6e-03	6.6e-03

scenario, an appropriate result given Table 2.2.

2.5.4 Results: Simulation 2

Simulation 2 introduced substantial data sparsity. Table 2.4 demonstrates the effect of positivity violations and near-violations on estimator behavior across 250 samples. The G-computation estimator remained minimally biased when the estimator $\bar{Q}_n$ was consistent; use of inconsistent $\bar{Q}_n$ resulted in bias. Given consistent estimators $\bar{Q}_n$ and g_n, the IPTW estimator was more biased than the other three estimators, as expected given the practical positivity violations present in the simulation. For this particular data-generating distribution and choice of misspecified model, misspecification of g_n increased the bias of the IPTW estimator further; however, this will not always be the case.

The finite sample performance of the A-IPTW and TMLE estimators was also affected by the presence of practical positivity violations. The DR estimators achieved the lowest MSE when 1) $\bar{Q}_n$ was consistent and 2) g_n was inconsistent but satisfied positivity (as a result either of truncation or of omission of W_2, a major source of positivity bias). Interestingly, in this simulation TMLE still did quite well when $\bar{Q}_n$ was inconsistent and the model used for g_n was correctly specified but its values bounded at $[0.025, 0.925]$.

Choice of bound imposed on g_n affected both the bias and variance of the IPTW, A-IPTW, and TMLE estimators. As expected, truncation of the IPTW weights improved the variance of the estimator but increased bias. Without additional diagnostic information, an analyst who observed the dramatic decline in the variance of the IPTW estimator that occurred with weight truncation might have concluded that truncation improved estimator performance; however, in this simulation weight truncation increased MSE. In contrast, and as predicted by theory, use of bounded values of g_n decreased MSE of the double robust estimators in spite of the inconsistency introduced to g_n.

Table 2.5 shows the mean and variance of the estimates of $ETA.Bias$ across 10 of the 250 samples. Based on the results shown in Table 2.4, a red flag diagnostic for the presence of bias due to positivity violations was needed for the IPTW estimator at all levels of bounding g_n, and for the TMLE estimator with unbounded g_n. (The A-IPTW estimator had a small

Table 2.3: True finite sample bias by specification (based on 250 samples of sample size 1000 with consistent g_n and Q_n) and mean and variance of estimated $ETA.Bias$ (based on the first 10 of the 250 samples) in Simulation 1: g_0 in [0.48,0.92], shown for unbounded g_n only.

		G-COMP	**IPTW**	**A-IPTW**	**TMLE**
	True finite sample bias	1.51e-03	5.95e-03	2.61e-04	-6.71e-06
Qcgc	Mean(ETA.Bias)	-4.21e-04	5.92e-04	-5.43e-04	-6.94e-04
	Variance(ETA.Bias)	2.23e-06	2.81e-06	2.34e-06	2.35e-06
	Mean(ETA.Bias)/True Bias	-2.79e-01	9.94e-02	-2.08e+00	1.03e+02
Qcgm	Mean(ETA.Bias)	6.17e-04	1.27e-03	4.17e-04	2.42e-04
	Variance(ETA.Bias)	7.32e-06	1.57e-05	6.48e-06	6.54e-06
	Mean(ETA.Bias)/True Bias	4.09e-01	2.14e-01	1.60e+00	-3.61e+01
Qmgc	Mean(ETA.Bias)	6.99e-04	1.51e-03	4.78e-04	3.05e-04
	Variance(ETA.Bias)	6.37e-06	8.18e-06	7.27e-06	7.25e-06
	Mean(ETA.Bias)/True Bias	4.63e-01	2.54e-01	1.83e+00	-4.54e+01

Table 2.4: Performance of estimators by specification and by bound on g_n in Simulation 2: g_0 in [0.001,1]. Results are based on 250 samples of size 1000.

Bound on g_n	**Qcgc**			**Qcgm**			**Qmgc**		
	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE
G-COMP									
None	0.007	0.009	0.009	0.007	0.009	0.009	1.145	0.025	1.336
[0.025,0.975]	0.007	0.009	0.009	0.007	0.009	0.009	1.145	0.025	1.336
[0.05,0.95]	0.007	0.009	0.009	0.007	0.009	0.009	1.145	0.025	1.336
[0.1,0.9]	0.007	0.009	0.009	0.007	0.009	0.009	1.145	0.025	1.336
IPTW									
None	0.544	0.693	0.989	1.547	0.267	2.660	0.544	0.693	0.989
[0.025,0.975]	1.080	0.090	1.257	1.807	0.077	3.340	1.080	0.090	1.257
[0.05,0.95]	1.437	0.059	2.123	2.062	0.054	4.306	1.437	0.059	2.123
[0.1,0.9]	1.935	0.043	3.787	2.456	0.043	6.076	1.935	0.043	3.787
A-IPTW									
None	0.080	0.966	0.972	-0.003	0.032	0.032	-0.096	16.978	16.987
[0.025,0.975]	0.012	0.017	0.017	0.006	0.017	0.017	0.430	0.035	0.219
[0.05,0.95]	0.011	0.014	0.014	0.009	0.014	0.014	0.556	0.025	0.334
[0.1,0.9]	0.009	0.011	0.011	0.008	0.011	0.011	0.706	0.020	0.519
TMLE									
None	0.251	0.478	0.540	0.026	0.059	0.060	-0.675	0.367	0.824
[0.025,0.975]	0.016	0.028	0.028	0.005	0.021	0.021	-0.004	0.049	0.049
[0.05,0.95]	0.013	0.019	0.020	0.010	0.016	0.017	0.163	0.027	0.054
[0.1,0.9]	0.010	0.014	0.014	0.009	0.013	0.013	0.384	0.018	0.166

to moderate level of bias with unbounded g_n; however the high variance of this estimator would have alerted an analyst to sparsity.) The parametric bootstrap correctly identified the presence of substantial $ETA.Bias$ in the IPTW estimator regardless of truncation level and in the TMLE estimator with unbounded g_n. It suggested minimal $ETA.Bias$ for the remaining estimators.

For correctly specified Q_n and g_n (g_n unbounded), the diagnostic captured 78% and 69% of the true finite sample bias of the IPTW and TMLE estimators, respectively. The fact that the true bias was underestimated in both cases illustrates a key limitation of the parametric bootstrap- its performance suffers when the target estimator is not asymptotically normally distributed (van der Vaart and Wellner [1996]). Bounding g_n improved the ability of the bootstrap to accurately diagnose bias by improving estimator behavior (in addition to adding a new source of bias due to use of inconsistent g_n). This finding suggests that practical application of the bootstrap to a given estimator should at minimum generate $ETA.Bias$ estimates for a single low level of bounding g_n in addition to any unbounded estimate. When g_n was bounded, the estimated $ETA.Bias$ for the IPTW estimator captured 96-98% of the true finite sample bias. The $ETA.Bias$ for the TMLE estimator with bounded g_n was accurately estimated to be minimal. As expected, misspecification of g_n or $\bar{Q}_n$ by excluding a key confounder lead to an estimated data generating distribution with less sparsity than the original, and as a result the parametric bootstrap underestimated the true extent of positivity bias for these model specifications.

While use of an unbounded g_n resulted in an underestimate of the true degree of $ETA.Bias$ for the IPTW and TMLE estimators, in this simulation the parametric bootstrap would still have functioned well as a diagnostic in each of the 10 samples considered. Tables 2.6 and 2.7 report the output that would have been available to an analyst applying the parametric bootstrap to the IPTW and TMLE estimators with unbounded g_n for each of the 10 samples. In all samples and for both estimators, the estimated $ETA.Bias$ was larger than the estimated standard error of the estimator, and was of significant magnitude relative to the point estimate of the causal effect.

Table 2.6 further demonstrates how the parametric bootstrap can be used to investigate the tradeoffs between bias due to weight truncation/bounding of g_n and positivity bias. The parametric bootstrap accurately diagnosed both an increase in the bias of the IPTW estimator with increasing truncation and a reduction in the bias of the TMLE estimator with truncation. When viewed in light of the standard error estimates under different levels of truncation, the diagnostic would have accurately suggested that truncation of g_n for the TMLE estimator was beneficial, while truncation of the weights for the IPTW estimator was of questionable benefit. (The parametric bootstrap can also be used to provide a more refined approach to choosing an optimal truncation constant based on estimated MSE.Bembom and van der Laan [2008])

Table 2.5: True finite sample bias and mean and variance of estimated $ETA.Bias$ (from first 10 of the 250 samples) by specification and bound on g_n, Simulation 2: g_0 in [0.001,1].

		Bound on g_n			
		None	[0.025,0.975]	[0.05,0.95]	[0.1,0.9]
G-COMP	True finite sample bias	7.01e-03	7.01e-03	7.01e-03	7.01e-03
Qcgc	Mean(ETA.Bias)	-8.51e-04	-8.51e-04	-8.51e-04	-8.51e-04
	Variance(ETA.Bias)	5.63e-06	5.63e-06	5.63e-06	5.63e-06
	Mean(ETA.Bias)/True bias	-1.21e-01	-1.21e-01	-1.21e-01	-1.21e-01
Qcgm	Mean(ETA.Bias)	2.39e-04	2.39e-04	2.39e-04	2.39e-04
	Variance(ETA.Bias)	1.37e-05	1.37e-05	1.37e-05	1.37e-05
	Mean(ETA.Bias)/True bias	3.41e-02	3.41e-02	3.41e-02	3.41e-02
Qmgc	Mean(ETA.Bias)	5.12e-04	5.12e-04	5.12e-04	5.12e-04
	Variance(ETA.Bias)	1.22e-05	1.22e-05	1.22e-05	1.22e-05
	Mean(ETA.Bias)/True bias	7.30e-02	7.30e-02	7.30e-02	7.30e-02
IPTW	True finite sample bias	5.44e-01	1.08e+00	1.44e+00	1.93e+00
Qcgc	Mean(ETA.Bias)	4.22e-01	1.04e+00	1.40e+00	1.90e+00
	Variance(ETA.Bias)	9.55e-03	2.19e-02	2.34e-02	2.39e-02
	Mean(ETA.Bias)/True Bias	7.76e-01	9.63e-01	9.73e-01	9.80e-01
Qcgm	Mean(ETA.Bias)	1.34e-01	4.83e-01	7.84e-01	1.23e+00
	Variance(ETA.Bias)	1.96e-03	1.08e-02	1.83e-02	2.40e-02
	Mean(ETA.Bias)/True Bias	2.46e-01	4.48e-01	5.46e-01	6.37e-01
Qmgc	Mean(ETA.Bias)	2.98e-01	7.39e-01	9.95e-01	1.35e+00
	Variance(ETA.Bias)	3.75e-03	9.65e-03	1.09e-02	1.36e-02
	Mean(ETA.Bias)/True Bias	5.48e-01	6.84e-01	6.93e-01	7.00e-01
A-IPTW	True finite sample bias	7.99e-02	1.25e-02	1.07e-02	8.78e-03
Qcgc	Mean(ETA.Bias)	1.86e-03	2.80e-03	5.89e-05	1.65e-03
	Variance(ETA.Bias)	1.51e-04	1.12e-05	4.68e-06	1.51e-05
	Mean(ETA.Bias)/True bias	2.32e-02	2.24e-01	5.50e-03	1.88e-01
Qcgm	Mean(ETA.Bias)	-3.68e-04	-6.36e-04	2.56e-05	5.72e-04
	Variance(ETA.Bias)	7.54e-05	1.16e-05	1.15e-05	1.53e-05
	Mean(ETA.Bias)/True bias	-4.60e-03	-5.09e-02	2.39e-03	6.51e-02
Qmgc	Mean(ETA.Bias)	-3.59e-04	1.21e-04	-1.18e-04	-1.09e-03
	Variance(ETA.Bias)	2.19e-04	1.04e-05	1.41e-05	5.31e-06
	Mean(ETA.Bias)/True bias	-4.50e-03	9.70e-03	-1.10e-02	-1.25e-01
TMLE	True finite sample bias	2.51e-01	1.60e-02	1.31e-02	9.98e-03
Qcgc	Mean(ETA.Bias)	1.74e-01	4.28e-03	2.65e-04	1.84e-03
	Variance(ETA.Bias)	3.26e-03	2.32e-05	6.26e-06	2.23e-05
	Mean(ETA.Bias)/True bias	6.94e-01	2.67e-01	2.02e-02	1.84e-01
Qcgm	Mean(ETA.Bias)	2.70e-02	-3.07e-04	2.15e-04	7.74e-04
	Variance(ETA.Bias)	2.88e-04	1.50e-05	1.27e-05	1.46e-05
	Mean(ETA.Bias)/True bias	1.08e-01	-1.92e-02	1.64e-02	7.76e-02
Qmgc	Mean(ETA.Bias)	1.11e-01	9.82e-04	-2.17e-04	-1.47e-03
	Variance(ETA.Bias)	8.95e-04	2.59e-05	2.52e-05	6.48e-06
	Mean(ETA.Bias)/True bias	4.44e-01	6.13e-02	-1.66e-02	-1.47e-01

Table 2.6: IPTW estimate, standared error and ETA.Bias estimate by sample and by bound on g_n with Qcgc, in Simulation 2: g_0 in [0.001,1]

	None			**[0.025,0.975]**			**[0.05,0.95]**			**[0.1,0.9]**		
	$\hat{\Psi}_{IPTW}$	SE	$ETA.Bias$	$\hat{\Psi}_{IPTW}$	SE	$ETA.Bias$	$\hat{\Psi}_{IPTW}$	SE	$ETA.Bias$	$\hat{\Psi}_{IPTW}$	SE	$ETA.Bias$
1	0.207	0.203	0.473	1.462	0.196	1.092	2.119	0.197	1.456	2.815	0.201	1.965
2	1.722	0.197	0.425	2.339	0.192	1.047	2.665	0.190	1.413	3.033	0.192	1.924
3	1.957	0.184	0.306	2.192	0.182	0.876	2.493	0.181	1.217	2.880	0.183	1.717
4	1.926	0.206	0.510	2.648	0.200	1.310	2.973	0.198	1.672	3.311	0.199	2.170
5	2.201	0.192	0.565	2.267	0.193	1.158	2.551	0.196	1.510	3.029	0.202	2.000
6	0.035	0.236	0.520	2.450	0.196	1.146	2.767	0.192	1.504	3.154	0.195	1.999
7	1.799	0.180	0.346	1.999	0.180	0.996	2.433	0.181	1.338	2.915	0.184	1.813
8	-0.471	0.215	0.420	1.938	0.193	1.007	2.400	0.194	1.398	2.978	0.196	1.922
9	2.749	0.184	0.391	2.769	0.185	0.977	2.828	0.186	1.326	3.088	0.189	1.822
10	-0.095	0.228	0.263	1.289	0.210	0.788	1.847	0.206	1.139	2.513	0.201	1.636
Mean	1.203	0.203	0.422	2.135	0.193	1.040	2.508	0.192	1.397	2.972	0.194	1.897

Table 2.7: TMLE estimate, standared error and ETA.Bias estimate by sample and by bound on g_n with Qcgc, in Simulation 2: g_0 in [0.001,1]

	None			**[0.025,0.975]**			**[0.05,0.95]**			**[0.1,0.9]**		
	$\hat{\Psi}_{TMLE}$	SE	$ETA.Bias$	$\hat{\Psi}_{TMLE}$	SE	$ETA.Bias$	$\hat{\Psi}_{TMLE}$	SE	$ETA.Bias$	$\hat{\Psi}_{TMLE}$	SE	$ETA.Bias$
1	0.827	0.197	0.172	0.982	0.105	0.001	0.975	0.077	0.001	0.965	0.063	0.003
2	0.734	0.114	0.153	1.094	0.100	-0.001	1.144	0.089	0.003	1.115	0.077	0.003
3	1.379	0.105	0.087	1.171	0.089	0.010	1.167	0.084	0.002	1.136	0.071	0.011
4	0.237	0.089	0.252	0.886	0.077	0.006	0.968	0.071	-0.001	1.016	0.071	-0.003
5	2.548	0.182	0.245	1.205	0.130	0.008	1.095	0.095	0.000	1.035	0.077	0.006
6	0.533	0.228	0.234	1.137	0.122	0.010	1.126	0.084	-0.001	1.083	0.071	0.000
7	1.781	0.184	0.150	1.143	0.138	0.002	1.159	0.095	0.001	1.128	0.077	0.004
8	1.066	0.114	0.188	0.950	0.095	0.003	0.919	0.084	-0.005	0.944	0.071	0.000
9	1.974	0.114	0.161	1.278	0.084	0.007	1.235	0.077	-0.001	1.176	0.071	-0.004
10	0.628	0.173	0.099	0.785	0.1451	-0.004	0.838	0.126	0.004	0.907	0.089	-0.003
Mean	1.171	0.170	0.174	1.063	0.114	0.004	1.063	0.095	0.000	1.051	0.071	0.002

to the analyst having an bias estimate due to misspecification of g_0. It's important to remind the reader that ETA.Bias includes bias both due to ETA and to bounding g_n.

We recommend that the parametric bootstrap be applied to the IPTW estimator in addition to the analyst's estimator of choice. Tables 2.5 and 2.6 illustrate the benefits of this approach. Diagnosis of substantial bias in the IPTW estimator due to positivity violations would have alerted an analyst that the G-computation estimator was relying heavily on extrapolation, and that the double robust estimators were sensitive to bias arising from misspecification of the model used to estimate $\bar{Q}_0$.

2.5.5 Results: Simulation 3

This simulation investigated the performance of the parametric bootstrap as a tool for diagnosing finite sample bias caused by collinearity between A and W, with the following objectives: 1) investigate further the utility of the parametric bootstrap in a setting in which estimators could not be assumed to be asymptotically normally distributed; 2) illustrate how use of a data-adaptive approach to fit Q_n can result in a poorly performing diagnostic tool unless specific measures are taken to ensure the bootstrapped data retains the sparsity present in the original data; and 3) investigate whether inclusion of the propensity score $g_n(1|W)$ as a covariate in Q_n improved the sensitivity of the diagnostic in the setting of collinearity.

Table 2.8: Performance of estimators by specification in Simulation 3: g_0 in [0.001,1], shown for unbounded g_n only.

	Qcgc			**Qdgd1**		
	Bias	Var	MSE	Bias	Var	MSE
G-COMP	0.133	0.038	0.055	0.212	0.027	0.072
IPTW	0.233	0.230	0.284	0.232	0.231	0.284
A-IPTW	0.134	0.038	0.055	0.175	0.027	0.057
TMLE	0.291	0.120	0.205	0.329	0.136	0.245

Table 2.8 demonstrates that all estimators exhibited substantial bias, even when $\bar{Q}_n$ and g_n were consistent. This remained true regardless of the level at which g_n was bounded; in the interest of space, results across bounding levels for g_n are not shown for this simulation. When stepwise selection was used to estimate Q_0, (forcing inclusion of A), the algorithm did not select $W2$ due to the collinearity with A. The consequences are reflected in the greater bias of $Qdgd1$ versus $Qcgc$ in those estimators that rely on Q_0.

Table 2.9: True finite sample bias for G-computation, IPTW and A-IPTW estimators and mean and variance of estimated $ETA.Bias$ (from first 10 of the 250 samples) by specification, Simulation 3: g_0 in [0.001,1], shown for unbounded g_n only.

G-COMP	True finite sample bias	1.33e-01
Qcgc	Mean(ETA.Bias)	4.18e-02
	Variance(ETA.Bias)	5.62e-03
	Mean(ETA.Bias)/True Bias	3.14e-01
Stepwise G-COMP	True finite sample bias	2.12e-01
Qdgd1	Mean(ETA.Bias)	1.97e-02
	Variance(ETA.Bias)	1.21e-03
	Mean(ETA.Bias)/True Bias	9.29e-02
Qdgd2	Mean(ETA.Bias)	1.17e-01
	Variance(ETA.Bias)	1.37e-02
	Mean(ETA.Bias)/True Bias	5.52e-01
IPTW	True finite sample bias	2.33e-01
Qcgc	Mean(ETA.Bias)	8.19e-02
	Variance(ETA.Bias)	4.89e-03
	Mean(ETA.Bias)/True Bias	3.51e-01
Stepwise IPTW	True finite sample bias	2.32e-01
Qdgd1	Mean(ETA.Bias)	7.03e-02
	Variance	5.44e-03
	Mean(ETA.Bias)/True Bias	3.03e-01
Qdgd2	Estimated ETA.Bias	1.41e-01
	Variance(ETA.Bias)	1.34e-02
	Mean(ETA.Bias)/True Bias	6.08e-01
A-IPTW	True finite sample bias	1.34e-01
Qcgc	Mean(ETA.Bias)	4.20e-02
	Variance(ETA.Bias)	5.63e-03
	Mean(ETA.Bias)/True Bias	3.14e-01
Stepwise A-IPTW	True finite sample bias	1.75e-01
Qdgd1	Mean(ETA.Bias)	1.47e-02
	Variance(ETA.Bias)	7.14e-04
	Mean(ETA.Bias)/True Bias	8.40e-02
Qdgd2	Mean(ETA.Bias)	9.66e-02
	Variance	1.22e-02
	Mean(ETA.Bias)/True Bias	5.52e-01

Table 2.10: True finite sample bias for TMLE estimators and mean and variance of estimated $ETA.Bias$ (from first 10 of the 250 samples) by specification, Simulation 3: g_0 in [0.001,1], shown for unbounded g_n only.

TMLE	True finite sample bias	2.91e-01
Qcgc	Mean(ETA.Bias)	1.70e-01
	Variance(ETA.Bias)	1.05e-02
	Mean(ETA.Bias)/True Bias	5.83e-01
Stepwise TMLE	True finite sample bias	3.29e-01
Qdgd1	Mean(ETA.Bias)	1.93e-01
	Variance(ETA.Bias)	1.24e-02
	Mean(ETA.Bias)/True Bias	5.87e-01
Qdgd2	Mean(ETA.Bias)	2.56e-01
	Variance(ETA.Bias)	1.53e-02
	Mean(ETA.Bias)/True Bias	7.78e-01

Table 2.11: IPTW estimate, standared error and ETA.Bias estimate by sample and by bound on g_n with Qcgc, in Simulation 3: g_0 in [0.001,1]

	None			**[0.025,0.975]**			**[0.05,0.95]**			**[0.1,0.9]**		
	$\hat{\Psi}_{IPTW}$	SE	$ETA.Bias$	$\hat{\Psi}_{IPTW}$	SE	$ETA.Bias$	$\hat{\Psi}_{IPTW}$	SE	$ETA.Bias$	$\hat{\Psi}_{IPTW}$	SE	$ETA.Bias$
1	0.549	0.109	0.156	0.611	0.104	0.258	0.610	0.088	0.261	0.533	0.068	0.239
2	0.589	0.052	0.083	0.594	0.052	0.160	0.534	0.052	0.144	0.421	0.050	0.123
3	0.095	0.156	0.087	0.391	0.120	0.159	0.454	0.090	0.167	0.460	0.065	0.148
4	0.740	0.050	0.090	0.751	0.050	0.067	0.761	0.050	0.026	0.670	0.051	-0.041
5	0.579	0.117	0.191	0.648	0.106	0.278	0.708	0.089	0.297	0.578	0.066	0.236
6	0.753	0.050	0.079	0.765	0.050	0.019	0.626	0.051	-0.078	0.494	0.051	-0.164
7	0.154	0.074	0.020	0.209	0.076	0.056	0.350	0.082	0.060	0.375	0.068	0.040
8	0.831	0.050	0.018	0.667	0.050	-0.058	0.526	0.050	-0.156	0.444	0.050	-0.224
9	1.147	0.048	-0.043	0.648	0.048	-0.102	0.557	0.048	-0.167	0.521	0.048	-0.230
10	0.043	0.154	0.138	0.492	0.109	0.238	0.592	0.081	0.243	0.538	0.062	0.214
Mean	0.548	0.096	0.082	0.578	0.082	0.107	0.572	0.070	0.080	0.503	0.059	0.034

Table 2.12: TMLE estimate, standared error and ETA.Bias estimate by sample and by bound on g_n with Qcgc, in Simulation 3: g_0 in [0.001,1]

	None			**[0.025,0.975]**			**[0.05,0.95]**			**[0.1,0.9]**		
	$\hat{\Psi}_{TMLE}$	SE	$ETA.Bias$	$\hat{\Psi}_{TMLE}$	SE	$ETA.Bias$	$\hat{\Psi}_{TMLE}$	SE	$ETA.Bias$	$\hat{\Psi}_{TMLE}$	SE	$ETA.Bias$
1	0.281	0.179	0.242	0.281	0.179	0.251	1.000	0.321	0.385	1.000	0.239	0.574
2	0.263	0.179	0.250	0.263	0.179	0.198	0.263	0.179	0.224	1.000	0.242	0.463
3	0.308	0.182	0.255	0.308	0.182	0.223	0.308	0.182	0.296	1.000	0.241	0.461
4	0.330	0.184	0.003	0.680	0.164	0.021	0.683	0.164	0.046	1.000	0.176	0.177
5	0.420	0.187	0.243	0.336	0.184	0.242	1.000	0.322	0.368	1.000	0.243	0.458
6	1.000	0.167	0.194	1.000	0.167	0.196	1.000	0.170	0.259	1.000	0.176	0.350
7	0.328	0.184	0.019	0.328	0.184	0.026	0.328	0.184	0.033	0.328	0.187	0.107
8	0.286	0.179	0.251	1.000	0.170	0.248	1.000	0.173	0.299	1.000	0.179	0.341
9	1.000	0.167	0.059	0.739	0.164	0.058	1.000	0.170	0.083	1.000	0.173	0.211
10	0.379	0.192	0.180	0.319	0.190	0.138	0.319	0.190	0.194	1.000	0.247	0.342
Mean	0.459	0.179	0.170	0.525	0.176	0.160	0.690	0.232	0.219	0.933	0.217	0.348

The parametric bootstrap underestimated $ETA.Bias$ more substantially in this simulation. It would have provided a reasonable albeit imperfect diagnostic tool. Tables 2.9 and 2.10 demonstrate that for all estimators, when $\bar{Q}_n$ and g_n were consistent the estimates of $ETA.Bias$ captured 30-35% of the true finite sample bias of the G-computation, IPTW, and A-IPTW estimators , and 58% of the finite sample bias of the TMLE estimator. Tables 2.11 and 2.12 show the sample-specific $ETA.Bias$ estimates for the IPTW and TMLE estimators. When compared by an analyst to the corresponding point and variance estimates for the target parameter, the diagnostic would have suggested caution in most but not all cases. Tables 2.9 and 2.10 further demonstrate that use of a stepwise algorithm that forces A to be included in $\bar{Q}_n$ generally resulted in a greater underestimate of $ETA.Bias$ because bootstrap data are simulated from a distribution in which sparsity plays less of a role. Retention of the propensity score in the fit of $\bar{Q}_0$ that was used to generate the bootstrap data ($Qdgd2$) improved the sensitivity of the diagnostic.

2.5.6 Discussion of Simulation Results

In summary, examination of the estimated treatment mechanism and corresponding propensity scores $g(a|W)$ may provide an initial alert to the presence of positivity violations; however, this approach does not provide a quantitative estimate of the resulting bias. The parametric bootstrap is a supplemental tool that allows the analyst to evaluate estimator behavior under a range of hypothetical data-generating distributions in which both the true value of the target parameter and the correct specification of nuisance parameter models is known. Further study of the performance of the diagnostic under a range of true and estimated data generating distributions is needed.

2.6 Data Example: HIV Resistance Mutations

2.6.1 Data and Question

We analyzed an observational cohort of HIV-infected patients in order to estimate the effect of mutations in the HIV protease enzyme on viral response to the antiretroviral drug lopinavir. The question, data, and analysis have been described previously.Bembom et al. [2009] Here, a simplified version of prior analyses was performed and the parametric bootstrap was applied to investigate the potential impact of positivity violations on results.

Briefly, baseline covariates, mutation profiles prior to treatment change, and viral response to therapy were collected for 401 treatment change episodes (TCEs) in which protease inhibitor-experienced subjects initiated a new antiretroviral regimen containing the drug lopinavir. We focused on 2 target mutations in the protease enzyme: p82AFST and p82MLC (present in 25% and 1% of TCEs, respectively). The data for each target mutation consisted of $O = (W, A, Y)$, where A was a binary indicator that the target mutation

was present prior to treatment change, W was a set of 35 baseline characteristics including summaries of past treatment history, mutations in the reverse transcriptase enzyme, and a genotypic susceptibility score for the background regimen (based on the Stanford scoring system; http://hivdb.stanford.edu/). The outcome Y was the change in $\log_{10}$(viral load) following initiation of the new antiretroviral regimen. The target observed data parameter was $E_W(E(Y|A=1,W)-E(Y|A=0,W))$, equal under (2.7) to the average treatment effect $E(Y_1-Y_0)$.

2.6.2 Methods

Effect estimates were obtained for each mutation using the IPTW estimator and TMLE with a logistic fluctuation. Gruber and van der Laan [2010b] $\bar{Q}_0$ and g_0 were estimated with stepwise forward selection of main terms based on the AIC criterion, using the `step` function in the *stats* v2.11.1 package in R. Estimators were implemented using both unbounded values for $g_n(A \mid W)$ and values truncated at $[0.025, 0.975]$. Following standard practice in much of the literature, standard errors were estimated using the influence curve, corresponding to the standard output for the `glm` and `tmle` functions in R, treating the values of g_n as fixed. The parametric bootstrap was used to estimate bias for each estimator using 1000 samples and the $ETA.Bias$ algorithm, with the `step` function rerun in each parametric bootstrap sample.

2.6.3 Results

Results for both mutations are presented in Table 2.13. p82AFST is known to be a major mutation for lopinavir resistance.Johnson et al. [2009] The current results support this finding; the IPTW and TMLE point estimates were similar and both suggested a significantly more positive change in viral load (corresponding to a less effective drug response) among subjects with the mutation as compared to those without it. The parametric bootstrap-based bias estimate was minimal, raising no red flag that these findings might be attributable to positivity bias.

The role of mutation p82CLM is less clear based on existing knowledge; depending on the scoring system used it is either not considered a lopinavir resistance mutation, or given an intermediate lopinavir resistance score (http://hivdb.stanford.edu/).Johnson et al. [2009] Initial inspection of the point estimates and standard errors in the current analysis would have suggested that p82CLM had a large and highly significant effect on lopinavir resistance. Application of the parametric bootstrap-based diagnostic, however, would have suggested that these results should be interpreted with caution. In particular, the bias estimate for the unbounded TMLE was larger than the estimated standard error, while the bias estimate for the unbounded IPTW estimator was of roughly the same magnitude. While neither bias estimate was of sufficient magnitude relative to the point estimate to change inference,

Table 2.13: Point estimate, standard error and parametric bootstrap-based bias estimates for the effect of two HIV resistance mutation on viral response, by estimator and bound on g_n.

	TMLE Estimator			**IPTW Estimator**		
	$\hat{\beta}_{TMLE}$	$\hat{SE}$	$ETA.Bias$	$\hat{\beta}_{IPTW}$	$\hat{SE}$	$ETA.Bias$
p82AFST						
$[0, 1]$	0.65	0.13	−0.01	0.66	0.15	−0.01
$[0.025, 0.975]$	0.62	0.13	0.00	0.66	0.15	−0.01
p82MLC						
$[0, 1]$	2.85	0.14	−0.37	1.29	0.14	0.09
$[0.025, 0.975]$	0.86	0.10	−0.01	0.80	0.23	0.08

their size relative to the corresponding standard errors would have suggested that further investigation was warranted.

In response, the non-parametric bootstrap (based on 1000 bootstrap samples) was applied to provide an alternative estimate of the standard error. Using this alternative approach, the standard errors for the unbounded TMLE and IPTW estimators of the effect of p82MLC were estimated to be 2.77 and 1.17, respectively. Non-parametric bootstrap-based standard error estimates for the bounded TMLE and IPTW estimators were lower (0.84 and 1.12, respectively), but still substantially higher than the initial naive standard error estimates. These revised standard error estimates dramatically changed interpretation of results, suggesting that the current analysis was unable to provide essentially any information on the presence, magnitude, or direction of the p82CLM effect. (Non-parametric bootstrap-based standard error estimates for p82AFST were also somewhat larger than initial estimates, but did not change inference).

In this example, $ETA.Bias$ is expected to include some non-positivity bias due to the curse of dimensionality. However, the resulting bias estimate should still be interpreted as highly optimistic (i.e. as an underestimate of the true finite sample bias). The parametric bootstrap sampled from estimates of g_0 and $\bar{Q}_0$ that had been fit using the `step` algorithm. This ensured that the estimators g_n and $\bar{Q}_n$ (which applied the same stepwise algorithm) would do a good job approximating $g_{\hat{P}_0}$ and $\bar{Q}_{\hat{P}_0}$ in each bootstrap sample. Clearly, no such guarantee exists for the true P_0. This simple example further illustrates the utility of the non-parametric bootstrap for standard error estimation in the setting of sparse data and positivity violations. In this particular example, the improved variance estimate provided by the non-parametric bootstrap was sufficient to prevent positivity violations from leading to incorrect inference. As demonstrated in the simulations, however, in other settings even accurate variance estimates may fail to alert the analyst to threats posed by positivity violations.

2.7 Practical Approaches for Estimation in the Presence of Positivity Violations

How should analysis proceed once threats to inference due to data sparsity have been identified? In this section we review several approaches to effect estimation in the presence of positivity violations. These include changing the projection function $h(a, V)$ used to defined the target parameter β, restricting the covariate adjustment set, restricting the sample, and redefining the causal effect of interest through the use of realistic and intention to treat parameters. Moore *et. al.* provide an extended review of these approaches. Moore et al. [2009] All four approaches can be viewed as a means to define a family of parameters that approximate the original target of inference to differing degrees. Estimators can then be defined that select among members of a given family based on the tradeoff between degree of divergence from the original target and identifiability.

2.7.1 Approach #1: Change the Projection Function $h(A, V)$

Throughout this paper we have focused on the target causal parameter $\beta(F_X, m, h)$ defined according to (2.5) as the projection of the $E_{F_X}(Y_a|V)$ on the marginal structural model $m(a, V|\beta)$. Choice of function $h(a, V)$ both defines the target parameter by specifying which values of (A, V) should be given greater weight when estimating β and, by assumption (2.9), defines the positivity assumption needed for β to be identifiable.

We have focused on parameters indexed by $h(a, V) = 1$, a choice that gives equal weight to estimating the counterfactual outcome for all values (a, v).Neugebauer and van der Laan [2007] Alternative choices of $h(a, V)$ can significantly weaken the needed positivity assumption. For example, if the target of inference only involves counterfactual outcomes among some restricted range $[c, d]$ of possible values $\mathcal{A}$, defining $h(a, V) = I(a \in [c, d])$ weakens the positivity assumption by requiring sufficient variability only in the assignment of treatment levels within the target range. In some settings, the causal parameter defined by such a projection over a limited range of $\mathcal{A}$ might be of substantial *a priori* interest. For example, one may wish to focus estimation of a drug dose response curve only on the range of doses considered reasonable for routine clinical use, rather than on the full range of doses theoretically possible or observed in a given data set.

An alternative approach, commonly employed in the context of IPTW estimation and introduced in Section 2.3.2, is to choose $h(a, V) = g(a|V)$, where $g(a|V) \equiv P(A = a|V)$ is the conditional probability of treatment given the covariates included in the marginal structural model. In the setting of IPTW estimation this choice corresponds to the use of stabilizing weights, a common approach to reducing both the variance of the IPTW estimator in the face of sparsity.Robins et al. [2000] When the target causal parameter is defined using a non-parametric marginal structural model, use of $h(a, V) = g(a, V)$ corresponds with a decision to define a target parameter that gives greater weight to those regions of the joint

distribution of (A, V) that are well-supported, and that relies on smoothing or extrapolation to a greater degree in areas that are not. Neugebauer and van der Laan [2007]

Use of a marginal structural working model makes clear that the utility of choosing $h(a, V) = g(a|V)$ as a method to approach data sparsity is not limited to the IPTW estimator. Recall that the G-computation estimator can be implemented by regressing predicted values for Y_a on (a, V) according to model $m(a, V|\beta)$ with weights provided by $h(a, V)$. When the projection function is chosen to be $g(a|V)$, this corresponds to a weighted regression in which weights are proportional to the degree of support in the data.

Even when one is ideally interested in the entire causal curve (implying a target parameter defined by choice $h(a, V) = 1$), specification of alternative choices for h offers a means of improving identifiability, at a cost of redefining the target parameter. For example, one can define a family of target parameters indexed by $h_\delta(a, V) = I(a \in [c(\delta), d(\delta)])$, where an increase in δ corresponds to progressive restriction on the range of treatment levels targeted by estimation. Fluctuation of δ thus corresponds to trading a focus on more limited areas of the causal curve for improved parameter identifiability. Selection of the final target from among this family can be based on an estimate of bias provided by the parametric bootstrap. For example, the bootstrap can be used to select the parameter with the smallest δ below some pre-specified threshold for allowable $ETA.Bias$.

2.7.2 Approach #2: Restrict the Adjustment Set

Exclusion of problematic W (i.e. those covariates resulting in positivity violations or near violations) from the adjustment set, provides a means to trade confounding bias for a reduction in positivity violations.Bembom et al. [2008] In some cases, exclusion of covariates from the adjustment set may come at little or no cost to bias in the estimate of the target parameter. In particular, a subset of W that excludes covariates responsible for positivity violations may still be sufficient to control for confounding. In other words, a subset $W' \subset W$ may exist for which both identifying assumptions (2.7) and (2.8) hold (i.e. $Y_a \coprod A \mid W'$ and $g_0(a|W') > 0, a \in \mathcal{A}$), while positivity fails for the full set of covariates. In practice, this approach can be implemented by first determining candidate subsets of W under which the positivity assumption holds, and then using causal graphs to assess whether any of these candidates is sufficient to control for confounding. Even when no such candidate set can be identified, background knowledge (or sensitivity analysis) may suggest that problematic W represent a minimal source of confounding bias (Moore et. al. provide an example).Moore et al. [2009] Often, however, those covariates that are most problematic from a positivity perspective are also strong confounders.

As suggested with respect to choice of projection function $h(a, V)$ in the previous section, the causal effect estimator can be fine-tuned to select the degree of restriction on the adjustment set W according to some pre-specified rule for eliminating covariates from the adjustment set, and the parametric bootstrap used to select the minimal degree of restriction that maintains $ETA.Bias$ below an acceptable threshold.Bembom et al. [2008] Also, the C-

TMLE estimator mentioned briefly in Section 2.3.3, which includes in the fit of g_n only those covariates that improve estimation of the target parameter, will restrict W in a "black-box" manner. In the case of substantial positivity violations, such approaches can result in small covariate adjustment sets. While such limited covariate adjustment accurately reflects a target parameter that is poorly supported by the available data, the resulting estimate can be difficult to interpret and will no longer carry a causal interpretation.

2.7.3 Approach # 3: Restrict the Sample

An alternative approach, sometimes referred to as "trimming", is to discard classes of subjects for whom there exists no or limited variability in observed treatment assignment. A causal effect is then estimated in the remaining subsample. This approach is popular in the econometrics and social science literature; Crump provides a recent review.Crump et al. [2006], LaLonde [1986], Heckman et al. [1997], Dehejia and S.Wahba [1999]

When the subset of covariates responsible for positivity violations is low or one dimensional, such an approach can be implemented simply by discarding subjects with covariate values not represented in all treatment groups. For example, say that one aims to estimate the average effect of a binary treatment, and in order to control for confounding needs to adjust for W, a covariate with possible levels $\{1, 2, 3, 4\}$. However, inspection of the data reveals that no one in the sample with $W = 4$ received treatment (ie. $g_n(1|W = 4) = 0$). The sample can be trimmed by excluding those subjects for whom $W = 4$ prior to applying a given causal effect estimator for the average treatment effect. As a result, the target parameter is shifted from $E(Y_1 - Y_0)$ to $E(Y_1 - Y_0|W < 4)$, and the positivity assumption (2.8) now holds (as $W = 4$ occurs with zero probability).

Often W is too high dimensional to make this straightforward implementation feasible; in such a case matching on the propensity score provides a means to trim the sample. There is an extensive literature on propensity score-based effect estimators; however such estimators are beyond the scope of the current review. Several potential problems arise with the use of trimming methods to address positivity violations. First, discarding subjects responsible for positivity violations shrinks sample size, and thus runs the risk of increasing the variance of the effect estimate. Further, sample size and the extent to which positivity violations arise by chance are closely related. Depending on how trimming is implemented, new positivity violations can be introduced as sample size shrinks. Second, restriction of the sample may result in a causal effect for a population of limited interest. In other words, as can occur with alternative approaches to improve identifiability by shifting the target of inference, the parameter actually estimated may be far from the initial target. Further, when the criterion used to restrict the sample involves a summary of high dimensional covariates, such as is provided the propensity score, it can be difficult to interpret the parameter estimated. Finally when treatment is longitudinal, the covariates responsible for positivity violations may themselves be affected by past treatment.Moore et al. [2009] Trimming to remove positivity violations in this setting amounts to conditioning on post-treatment covariates and can thus

introduce new bias.

Crump proposes an approach to trimming that falls within the general strategy of redefining the target parameter in order to explicitly capture the tradeoff between parameter identifiability and proximity to the initial target.Crump et al. [2006] In addition to focusing on the treatment effect in an *a priori* specified target population, he defines an alternative target parameter corresponding to the average treatment effect in that subsample of the population for which the most precise estimate can be achieved. Crump further suggests the potential for extending this approach to achieve an optimal (according to some user-specified criteria) tradeoff between the representativeness of the subsample in which the effect is estimated and the variance of the estimate.

2.7.4 Approach #4: Change the Intervention of Interest

A final alternative for improving the identifiability of a causal parameter in the presence of positivity violations is to redefine the intervention of interest. Realistic rules rely on an estimate of the propensity score $g(a|W)$ to define interventions that explicitly avoid positivity violations. This ensures that the causal parameter estimated is sufficiently supported by existing data.

Realistic interventions avoid positivity violations by first identifying subjects for whom a given treatment assignment is not realistic (i.e. subjects whose propensity score for a given treatment is small or zero) and then assigning an alternative treatment with better data support to those individuals. Such an approach is made possible by focusing on the causal effects of dynamic treatment regimes.van der Laan and Petersen [2007], Robins et al. [2008] The causal parameters described thus far are summaries of the counterfactual outcome distribution under a fixed treatment applied uniformly across the target population. In contrast, a dynamic regime assigns treatment in response to patient covariate values. This characteristic makes it possible to define interventions under which a subject is only assigned treatments that are possible (or "realistic") given a subject's covariate values.

To continue the previous example in which no subjects with $W = 4$ were treated, a realistic treatment rule might take the form "treat only those subjects with W less than 4." More formally, let $d(W)$ refer to a treatment rule that deterministically assigns a treatment $a \in \mathcal{A}$ based on a subject's covariates W and consider the rule $d(W) = I(W < 4)$. Let Y_d denote the counterfactual outcome under the treatment rule $d(W)$, which corresponds to treating a subject if and only if his or her covariate W is below 4. In this example $E(Y_0)$ is identified as $\sum_w E(Y|W = w, A = 0)P(W = w)$; however, since $E(Y|W = w, A = 1)$ is undefined for $W = 4$, $E(Y_1)$ is not identified (unless we are willing to extrapolate based on $W < 4$). In contrast, $E(Y_d)$ is identified by the non-parametric G-computation formula: $\sum_w E(Y = y|W = w, A = d(W))P(W = w)$. Thus the average treatment effect $E(Y_d - Y_0)$, but not $E(Y_1 - Y_0)$, is identified. The redefined causal parameter can be interpreted as the difference in expected counterfactual outcome if only those subjects with $W < 4$ were treated as compared to the outcome if no one were treated.

More generally, realistic rules indexed by a given static treatment a assign a only to those individuals for whom the probability of receiving a is greater than some user-specified probability α (such as $\alpha > 0.05$). Let $d(a, W)$ denote the rule indexed by static treatment a. If A is binary, then $d(1, W) = 1$ if $g(1|W) > \alpha$, otherwise $d(1, W) = 0$. Similarly, $d(0, W) = 0$ if $g(0|W) > \alpha$; otherwise $d(0, W) = 1$. Realistic causal parameters are defined as some parameter of the distribution of $Y_{d(a,W)}$ (possibly conditional on some subset of baseline covariates $V \subset W$). Estimation of the causal effects of dynamic rules $d(W)$ allows the positivity assumption to be relaxed to $g(d(W)|W) > 0$ -a.e (i.e. only those treatments that would be assigned based on rule d to patients with covariates W need to occur with positive probability within strata of W). Realistic rules $d(a, W)$ are designed to satisfy this assumption by definition.

When a given treatment level a is unrealistic (i.e. when $g(a \mid W) < \alpha$), realistic rules assign an alternative from among viable (well-supported) choices. Choice of an alternative is straightforward when treatment is binary. When treatment has more than two levels, however, a rule for selecting the alternative treatment level is needed. One option is to assign a treatment level that is as close as possible to the orignal assignment while still remaining realistic. For example, if high doses of drugs occur with low probability in a certain subset of the population, a realistic rule might assign the maximum dose that occurs with probability $> \alpha$ in that subset. An alternative class of dynamic regimes, referred to as "intent-to-treat" rules, instead assign a subject to his or her observed treatment value if an initial assignment is deemed unrealistic. Moore, *et. al.* and Bembom, *et. al.* provide illustrations of both of these types of realistic rules using simulated and real data.Moore et al. [2009], Bembom and van der Laan [2007]

The causal effects of realistic rules clearly differ from their static counterparts. The extent to which the new target parameter diverges from the initial parameter of interest depends on both the extent to which positivity violations occur in the finite sample (i.e. the extent of support available in the data for the initial target parameter) and on a user-supplied threshold α. The parametric bootstrap approach presented in Section 2.4 can be employed to data-adaptively select α based on the level of $ETA.Bias$ deemed acceptable.Bembom and van der Laan [2007]

2.7.5 Selection Among a Family of Parameters

Each of the methods described for estimating causal effects in the presence of data sparsity corresponds to a particular strategy for altering the target parameter in exchange for improved identifiability. In each case, we have outlined how this tradeoff could be made systematically, based on some user-specified criterion such as the bias estimate provided by the parametric bootstrap. We now summarize this general approach in terms of a formal method for estimation in the face of positivity violations.

1. Define a family of parameters. The family should include the initial target of inference

together with a set of related parameters, indexed by γ in index set I, where γ represents the extent to which a given family member trades improved identifiability for decreased proximity to the initial target. In the examples given in the previous section, γ could be used to index a set of projection functions $h(a, V)$ based on an increasingly restrictive range of the possible values $\mathcal{A}$, degree to which the adjustment covariate set or sample is restricted, or choice of a threshold for defining a realistic rule.

2. Apply the parametric bootstrap to generate an estimate $ETA.Bias$ for each $\gamma \in I$. In particular, this involves estimating the data generating distribution, simulating new data from this estimate, and then applying an estimator to each target indexed by γ.

3. Select the target parameter from among the set that fall below a pre-specified threshold for acceptable $ETA.Bias$. In particular, select the parameter from within this set that is indexed by the value γ that corresponds to the greatest proximity to the initial target.

This approach allows an estimator to be defined in terms of an algorithm that identifies and estimates the parameter within a candidate family that is as close to the initial target of inference as possible while remaining within some user-supplied limit on the extent of tolerable positivity violations.

2.8 Conclusions

The identifiability of causal effects relies on sufficient variation in treatment assignment within covariate strata. The strong version of positivity requires that each possible treatment occur with positive probability in each covariate strata; depending on the model and target parameter, this assumption can be relaxed to some extent. In addition to assessing identifiability based on measurement of and control for sufficient confounders, data analyses should directly assess threats to identifiability based on positivity violations. The parametric bootstrap is a practical tool for assessing such threats, and provides a quantitative estimator-specific estimate of bias arising due to positivity violations.

The objective of the parametric bootstrap diagnostic is to raise a red flag in settings where positivity violations (as well as bounding of g_n) may be resulting in bias of sufficient magnitude to threaten reliable inference. The simulations showed that the diagnostic worked best when (1) Q_n and g_n were consistently estimated; (2) g_n was at least minimally bounded so that the estimator was more likely to be asymptotically normal; and, (3) any data-adaptive algorithm used to fit $\bar{Q}_0$ was forced to include not only A but also the propensity score in order to retain sparsity in the bootstrapped distribution. Although the diagnostic may underestimate the true $ETA.Bias$, in the simulations presented here the diagnostic was generally successful in raising a red-flag for bias due to positivity violations in the settings

where such a warning was needed. The performance of the diagnostic should be further investigated under a range of true and estimated data generating distributions, however.

This paper has focused on the positivity assumption for the causal effect of a treatment assigned at a single time point. Extension to a longitudinal setting in which the goal is to estimate the effect of multiple treatments assigned sequentially over time introduces considerable additional complexity. First, practical violations of the positivity assumption can arise more readily in this setting. Under the longitudinal version of the positivity assumption the conditional probability of each possible treatment history should remain positive regardless of covariate history. However, this probability is the product of time point-specific treatment probabilities given the past. When the product is taken over multiple time points it is easy for treatment histories with very small conditional probabilities to arise. Second, longitudinal data make it harder to diagnose the bias arising due to positivity violations. Implementation of the parametric bootstrap in longitudinal settings requires Monte Carlo simulation both to implement the G-computation estimator and to generate each bootstrap sample. In particular, this requires estimating and sampling from the time-point specific conditional distributions of all covariates and treatment given the past. Additional research on assessing the impact of of positivity bias on longitudinal causal parameters is needed, including investigation of the parametric bootstrap in this setting.

When positivity violations occur for structural reasons rather than due to chance, a causal parameter that avoids these positivity violations will often be of substantial interest. For example, when certain treatment levels are contraindicated for certain types of individuals, the average treatment effect in the population may be of less interest than the effect of treatment among that subset of the population without contraindications, or alternatively, the effect of an intervention that assigns treatment only to those subjects without contraindications. Similarly, the effect of a multilevel treatment may be of greatest interest for only a subset of treatment levels.

In other cases researchers may be happy to settle for a better estimate of a less interesting parameter. Sample restriction, estimation of realistic parameters, and change in projection function $h(a, V)$ all change the causal effect being estimated; in contrast, restriction of the covariate adjustment set often results in estimation of a non-causal parameter. However, all of these approaches can be understood as means to shift from a poorly identified initial target towards a parameter that is less ambitious but more fully supported by the available data. The new estimand is not determined *a priori* by the question of interest, but rather is driven by the observed data distribution in the finite sample at hand. There is thus an explicit tradeoff between identifiability and proximity to the initial target of inference. Ideally, this tradeoff will be made in a systematic way rather than on an *ad hoc* basis at the discretion of the investigator. Definition of an estimator that selects among a family of parameters according to some pre-specified criteria is a means to formalize this tradeoff. An estimate of bias based on the parametric bootstrap can be used to implement the tradeoff in practice.

The parametric bootstrap also provides a means to optimize estimator performance without changing the target parameter. The parametric bootstrap provides an estimate of the

whole sampling distribution of a candidate estimator, and thus can be used to estimate MSE and fine-tune estimator performance based on this estimate. Bembom *et. al.* illustrate this approach by using the bootstrap to data-adpatively select the level of weight truncation that minimizes the estimated MSE of the IPTW estimator; the same method can also be used to minimize estimated MSE using alternative approaches such as progressive restriction of the adjustment set. We emphasize, however, that use of the parametric bootstrap to minimize estimator MSE is fundamentally different than use of the parametric bootstrap to select among a family of parameters, as described in Section 2.7.5. The former represents a means of improving estimator performance for the same target parameter (by fine-tuning the estimator to optimize bias-variance tradeoff). In contrast, the family of parameters approach shifts the target of inference to a parameter that is adequately supported by the data.

In summary, we offer the following advice for applied analyses: First, define the causal effect of interest based on careful consideration of structural positivity violations. Second, consider estimator behavior in the context of positivity violations when selecting an estimator. Third, apply the parametric bootstrap to quantify the extent of estimator bias under data simulated to approximate the true data generating distribution. Fourth, when positivity violations are a concern, choose an estimator that selects systematically among a family of parameters based on the tradeoff between data support and proximity to the initial target of inference.

Chapter 3

The Relative Performance of Targeted Maximum Likelihood Estimators

3.1 Introduction

This chapter delves more deeply into the relative performance of different estimators under violations of the positivity assumption, with particular emphasis on a variety of double robust (DR) estimators. The chapter mostly replicates a paper that is currently under review for publication in the International Journal of Biostatistics, with title identical to the chapter title and coauthored by Susan Gruber (co-first author), Mark van der laan and Jasjeet S. Sekhon. Much of the material is also introduced in Rose and van der Laan [Eds.]. The paper was motivated by recent literature in which there is much debate on the relative performance of DR estimators when the positivity assumption is violated. In particular, Kang and Schafer [2007] (KS) demonstrate the fragility of DR estimators in a simulation study with near, or practical, positivity violations. They focus on a simple missing data problem in which one wishes to estimate the mean of an outcome that is subject to missingness and all possible covariates for predicting missingness are measured. Responses by Robins et al. [2007], Tsiatis and Davidian [2007], Tan [2007] and Ridgeway and McCaffrey [2007] further explore the challenges faced by DR estimators and offer suggestions for improving their stability.

In their article, KS introduce a variety of DR estimators and compare them to non-DR inverse probability of censoring (IPCW) estimators, as well as to a simple parametric model based ordinary least squares (OLS) estimator. As the KS simulation has practical positivity violations, some values of both the true and estimated missingness mechanism are very close to zero. In this situation, the IPCW will be extremely large for some observations of the sample. Therefore, DR and non-DR estimators that rely on IPCW may be unreliable. As a result, KS warn against the routine use of estimators that rely on IPCW, including DR estimators: this is in agreement with other literature analyzing the issue (Robins [1986, 1987a, 1999], Robins and Wang [2000], van der Laan and Robins [2003]), showing simulations

demonstrating the extreme sparsity bias of IPCW-estimators (e.g., Neugebauer and van der Laan [2005]), diagnosing violations of the positivity assumptions in response to this concern (Petersen et al. [2010], Wang et al. [2006a], Moore et al. [2009], Cole and Hernan [2008], Kish [1992], Bembom and van der Laan [2008]), data adaptive selection of the truncation constant to control the influence of weighting (Bembom and van der Laan [2008], and selecting parameters that are relying on realistic assumptions (see van der Laan and Petersen [2007], and Petersen et al. [2010]).

The particular simulation in KS also gives rise to a situation in which under dual misspecification, the OLS estimator outperforms all of the presented DR estimators. While this is an interesting issue, it is not the main focus here. In our view, dual misspecification brings up the need for other strategies for improving the robustness of estimators in general, such as incorporating data adaptive estimation instead of relying on parametric regression models for the missingness mechanism and the conditional distribution of responses, an idea echoed in the responses by Tsiatis and Davidian [2007] and Ridgeway and McCaffrey [2007], and standardly incorporated in the UC Berkeley literature on targeted maximum likelihood estimation (e.g., van der Laan and Rubin [2006], van der Laan et al. [2009]). In particular, we note that a statistical estimation problem is also defined by the statistical model, which, in this case, is defined by a nonparametric model: such models require data adaptive estimators in order to claim that the estimator is consistent. Nonetheless, we explicitly demonstrate the impact of the utilization of machine learning on the simulation results in a final section of this article.

In their response to the KS paper, Robins et al. [2007] point out that a desirable property of DR estimators is "boundedness," in that for a finite sample, estimators of the mean response fall in the parameter space with probability 1. Estimators that impose such a restriction can introduce new bias but avoid the challenges of highly variable weights. Robins et al. [2007] discuss ways in which to guarantee that "boundedness" holds and present two classes of bounded estimators–regression DR estimators and bounded Horvitz-Thompson DR estimators. We define examples of these estimators below, and we evaluate their relative performance in Section 3.6. The response by Tsiatis and Davidian [2007] offers strategies for constructing estimators that are more robust under the circumstances in the KS simulations. In particular, to address positivity violations, they suggest an estimator that uses IPCW only for observations with missingness mechanism values that are not close to zero, while using regression predictions for the observations with very small missingness mechanism values. One might consider either a hard cutoff for dividing observations or weighting each part of the influence curve by the estimated missingness mechanism. Tan [2007] also points to an improved locally efficient double robust estimator (Tan [2006]) that is able to maintain double robustness as well as provides guaranteed improvement relative to an initial estimator, improving on such type of estimators that had an algebraic similar form but failed to guarantee both properties (Robins et al. [1994], and see also van der Laan and Robins [2003]). Many responders also make valuable suggestions regarding the dual misspecification challenge.

In the current paper, we add targeted maximum likelihood estimators (TMLE's), or more generally, targeted minimum loss based estimators (van der Laan and Rubin [2006]) to the debate on the relative performance of DR estimators under practical violations of the positivity assumption in the particular simple missing data problem set forth by KS. TMLE's involve a two-step procedure in which one first estimates the conditional expectation of the outcome, given the covariates, and then updates this initial estimator, targeting bias reduction of the parameter of interest, rather than the overall conditional mean of the outcome given the covariates. The second step requires specification of a loss-function (e.g., log-likelihood loss function) and a parametric submodel through the initial regression, so that one can fit the parametric sub-model by minimizing the empirical risk (e.g., maximizing the log-likelihood). The estimator of the target parameter is then defined as the corresponding substitution estimator. Because TMLE's are substitution estimators, they not only respect the global bounds of the parameter and data (and thus satisfy the "boundedness" property defined by Robins et al. [2007]), but, even more importantly, they respect the fact that the true parameter value is a particular function of the data generating probability distribution.

TMLE's are double robust and asymptotically efficient. Moreover, TMLE's can incorporate data-adaptive likelihood or loss based estimation procedures to estimate both the conditional expectation of the outcome and the missingness mechanism.The TMLE also allows the incorporation of targeted estimation of the censoring/treatment mechanism, as embodied by the collaborative TMLE (C-TMLE), thereby fully confronting a long standing problem of how to select covariates in the propensity score/missingness mechanism of DR-estimators. In this article, we compare the performance of TMLE's to other DR estimators in the literature using the exact simulation study presented in the KS paper. We also make slight modifications to the KS simulation, in order to make the estimation even more challenging.

The DR parametric regression estimator of Scharfstein et al. [1999]), which was included in the response of Robins et al. [2007], is a particular special case of a TMLE (Rosenblum and van der Laan [2010b]). It defines a clever parametric initial regression for which the update step of the general TMLE-algorithm introduced in van der Laan and Rubin [2006] results in a zero-update, and is thus not needed. Such a TMLE falls in the class of TMLE's defined by an initial regression estimator, a squared error loss function and univariate linear regression submodel (coding the fluctuations of the initial regression estimator for the TMLE-update step). Such TMLE's for continuous outcomes (contrary to the excellent robustness of the TMLE for binary outcome based on the log-likelihood loss function and logistic regression submodel) suffer from great sensitivity to violations of the positivity assumptions, as was also observed in the simulations presented in the Kang and Shafer debate. As explained in (Gruber and van der Laan [2010a]) the problem with this TMLE defined by the squared error loss function and univariate linear regression submodel is that its updates are not subject to any bounds implied by the statistical model or data: that is, it is not using a parametric *sub*-model, an important principle of the general TMLE algorithm. A valid TMLE for continuous outcomes, defined by a different loss function and a univariate logistic regression parametric submodel,

has been recently presented (Gruber and van der Laan [2010a]), which demonstrates that the previously observed sensitivity of these two estimators to the positivity assumption was due to those specific choices.

The remainder of this chapter is organized as follows. Section 3.2 presents notation, which deviates from that presented in KS, for the data structure and parameter of interest. Section 3.3 formally defines the positivity assumption and gives an overview of causes, diagnostics and responses to violations. Section 3.4 defines the estimators on which we focus in this paper, including a sample of estimators in the literature and TMLE's. Section 3.5 then compares the performance of the estimators in the original and modified KS simulation. Finally, Section 3.6 summarizes the results and Section 3.8 concludes with a discussion of the findings.

3.2 Data Structure, Statistical Model, and Parameter of Interest

Consider an observed data set consisting of n independent and identically distributed (i.i.d) observations of $O = (W, \Delta, \Delta Y) \sim P_0$. W is a vector of covariates, and $\Delta = 1$ indicates whether Y, a continuous outcome, is observed. P_0 denotes the true distribution of O, from which all observations are sampled. We view O as a missing data structure on a hypothetical full data structure $X = (W, Y)$, which contains the true, or potential, value of Y for all observations, as if no values are missing. We assume Y is missing at random (MAR) such that $P_0(\Delta = 1 \mid X) = g_0(1 \mid W)$. In other words, we assume there are no unobserved confounders of the relationship between missingness Δ and the outcome Y.

We define $Q_0 = \{Q_{0,W}, \bar{Q}_0\}$, where $Q_{0,W}(w) \equiv P_0(W = w)$ and $\bar{Q}_0(W) \equiv E_0(Y \mid \Delta = 1, W)$. We make no assumptions about Q_0. The generalized Cramer-Rao information bound for any parameter of Q_0 does not depend on the statistical model for the missingness mechanism g_0. The parameter of interest is the mean outcome $E_0 Y$ for the sampled population, as if there were not missing observations of Y. Due to the MAR assumption and the positivity assumption defined below, our target parameter is identified from P_0 by the following mapping from Q_0:

$$\mu(P_0) = E_0(Y) = E_0(\bar{Q}_0(W)).$$

3.3 The Positivity Assumption

The identifiability of the parameter of interest $\mu(P_0)$ requires MAR and adequate support in the data. Regarding the latter, it requires that within each stratum of W, there is positive probability that Y is not missing. This requirement is often referred to as the positivity assumption. Formally, for our target parameter, the positivity assumption requires that:

$$g_0(\Delta = 1 \mid W) > 0 \ P_0\text{-almost everywhere.} \tag{3.1}$$

The positivity assumption is specific to the the target parameter. For example, the positivity assumption of the target parameter $E_0\{E_0(Y \mid A = 1, W) - E_0(Y \mid A = 0, W)\}$ of the probability distribution of $O = (W, A, Y)$, representing the additive causal effect under causal assumptions, requires that within each stratum there is a positive probability for all possible treatment assignments. For example, if A is a binary treatment, then positivity requires that $0 < g_0(A = 1|W) < 1$. (The assumption is often referred to as the experimental treatment assignment (ETA) assumption for causal parameters.) In addition to being parameter-specific, the positivity assumption is also model-specific. Parametric model assumptions, which extrapolate to regions of the joint distribution of (A,W) that may not be supported in the data, allow for weakening the positivity assumption (Petersen et al. [2010]). However, analysts need to be sure that their parametric assumptions actually hold true, which may be difficult if not impossible.

Violations and near violations of the positivity assumption can arise for two reasons. First, it may be theoretically impossible or highly unlikely for the outcome Y to be observed for certain covariate values in the population of interest. The threat to identifiability due to such structural violations of positivity exists regardless of the sample size. Second, given a finite sample, the probability of the outcome being observed for some covariate values might be so small that the observed sample cannot be distinguished from a sample drawn under a theoretical violation of the positivity assumption. The effect of such practical violations of the positivity assumption are sample size specific, and the resulting sparse data bias and inflated variance are often as dramatic as under structural violations.

Several approaches for diagnosing bias due to positivity violations have been suggested (see Petersen et al. [2010] for an overview). Analysts may assess the distribution of Δ within covariate strata (or in the case of causal parameters, the distribution of treatment assignment), but this method is not practical with high dimensional covariate sets or with continuous or multi-level covariates, and also provides no quantitative measure of the resulting sparse-data bias. Analysts may also assess the distribution of the estimated missingness mechanism scores, $g_n(\Delta = 1|W)$, or inverse probability weights. While this approach may indicate positivity violations, it does not provide any information on the extent of potential bias of the chosen estimator. Wang et al. [2006b] introduce and Petersen et al. [2010] further discuss a diagnostic that provides an estimate of positivity bias for any candidate estimator, which is based on a parametric bootstrap. Bias estimates of similar or larger magnitude than an estimate's standard error can raise a red flag to analysts that inference for their target parameter is threatened by lack of positivity.

When censoring probabilities are close to 0 (or 1 in the case of an effect parameter), a common practice is to truncate the probabilities or the resulting inverse probability weights, either at fixed levels or at percentiles (Petersen et al. [2010], Wang et al. [2006a], Moore et al. [2009], Cole and Hernan [2008], Kish [1992], Bembom and van der Laan [2008]). The

practice limits the influence of observations with large unbounded weights, which may reduce positivity bias and rein in inflated variance. However, this practice may also introduce bias, due to misspecification of the missingness mechanism g_n. The extent to which truncating g_n hurts or helps the performance of an estimator depends on the level of truncation, the estimator and the distribution of the data. In our simulations below, we examine the effect of truncating missingness probabilities for all estimators that we introduce in the next section.

3.4 Estimators of a Mean Outcome when the Outcome is Subject to Missingness

3.4.1 Estimators in the Literature

As a benchmark, KS compare all estimators in their paper to the ordinary least squares (OLS) estimator. For the target parameter, the OLS estimator is equivalent to the G-computation estimator based on a linear regression model. It is defined as:

$$\mu_{n,OLS} = \frac{1}{n}\sum_{i=1}^{n} \bar{Q}_n^0(W_i).$$

where $\bar{Q}_n^0 = m_{\beta_n}$ is an initial linear regression fit of $\bar{Q}_0$, and β_n is given by:

$$\beta_n = \arg\min_{\beta} \sum_{i=1}^{n} \Delta_i (Y_i - m_\beta(W_i))^2.$$

Under violation of the positivity assumption, the OLS estimator, when defined, extrapolates from strata of W in which there is support to strata of W that lack adequate support. The extrapolation depends on the validity of the linear regression model, and misspecification leads to bias.

KS present comparisons of several DR (and non-DR) estimators. We focus on just a couple of them here. Using our terminology with the terminology and abbreviations from KS in parenthesis the estimators we compare are: the weighted least squares (WLS) estimator (regression estimation with inverse-propensity weighted coefficients, $\hat{\mu}_{WLS}$) and the augmented IPCW (A-IPCW) estimator (regression estimation with residual bias correction, $\hat{\mu}_{BC-OLS}$). Both of these DR estimators are defined below.

The WLS estimator is defined as:

$$\mu_{n,WLS} = \frac{1}{n}\sum_{i=1}^{n} m_{\beta_n}(W_i),$$

where

$$\beta_n = \arg\min_{\beta} \sum_{i=1}^{n} \frac{\Delta_i}{g_n(1|W_i)}(Y_i - m_\beta(W_i))^2.$$

The A-IPCW estimator, introduced by J.M. Robins and Zhao [1994], is then defined as:

$$\mu_{n,A-IPCW} = \bar{Q}_n^0(W_i) + \frac{1}{n}\sum_i \frac{\Delta_i}{g_n(1|W_i)}(Y_i - \bar{Q}_n^0(W_i)).$$

Both of these estimators rely on estimators of $\bar{Q}_0$ and g_0. They are consistent if $\bar{Q}_n^0$ or g_n is consistent, and efficient if both are consistent. Under positivity violations, however, these estimators rely on the consistency of $\bar{Q}_n^0$, and require that g_n converges to a limit that satisfies the positivity assumption (see e.g., van der Laan and Robins [2003]).

Additionally, in comments on KS, Robins et al. [2007] introduce bounded Horvitz-Thompson (BHT) estimators, which, as the name suggests, are bounded, in that for finite sample sizes the estimates are guaranteed to fall in the parameter space. A BHT estimator is defined as:

$$\mu_{n,BHT} = \bar{Q}_n^0(W) + \frac{1}{n}\sum_i \frac{\Delta_i}{g_{n_{EXT}}(1|W_i)}(Y_i - \bar{Q}_n^0(W_i)).$$

This is equivalent to the A-IPTW estimator, but estimating $g_0(1 \mid W)$ by fitting the following logistic regression model

$$\text{logit}P_{EXT}(\Delta = 1|W) = \alpha^T W + \phi h_n(W),$$

and $h_n(W) = \bar{Q}_n^0(W) - \frac{1}{n}\sum_{i=1}^n \bar{Q}_n^0(W_i)$.

We also include another important class of doubly robust, locally efficient, regression-based estimators introduced by Scharfstein et al. [1999], further discussed in Robins [1999] and compared to the TMLE's as defined in this paper in Rosenblum and van der Laan [2010b]. This estimator is based on a parametric regression model which includes a "clever covariate" that incorporates inverse probability weights. We use the abbreviation PRC. The estimator is defined as:

$$\psi_{n,PRC} = \frac{1}{n}\sum_{i=1}^{n} \bar{Q}_n^{'}(W_i),$$

where $Q_n^{'}(W) = m_{\beta_n,\epsilon_n}(W)$ and $m_{\beta,\epsilon}(W)$ is a parametric model, which includes the clever covariate $H_{g_n}^*(W) = \frac{1}{g_n(1|W)}$, and (β_n, ϵ_n) is the OLS.

3.4.2 TMLE's

We compare the above estimators with several versions of TMLE's. The targeted maximum likelihood procedure was first introduced in van der Laan and Rubin [2006]. For a compilation of current and past work on targeted maximum likelihood estimation, see van der Laan et al. [2009].

In contrast to the estimating equation-based DR estimators defined above (WLS, A-IPCW and BHT), the PRC estimator and TMLE's are DR *substitution* estimators. TMLE's are based on an update of an initial estimator of P_0 that fluctuates the fit with a fit of a clever parametric submodel. Assuming a valid parametric submodel is selected, TMLE's do not only respect the bounds on the outcome implied by the statistical model or data, but also respect that the true target parameter value is a specified function of the data generating distribution. Due to respecting this information, the TMLE does not only respect the local bounds of the statistical model by being asymptotically (locally) efficient (as the other DR estimators), but also respect the global constraints of the statistical model. Being a substitution estimator is particularly important under sparsity, as implied by violations of the positivity assumptions.

Although our target parameter involves a continuous Y, to introduce the TMLE for the mean outcome, we begin by defining the TMLE for a binary Y. In this case, the TMLE is defined as:

$$\mu_{n,TMLE} = \frac{1}{n}\sum_{i=1}^{n} \bar{Q}_n^*(W_i), \tag{3.2}$$

where we use the logistic regression submodel

$$\text{logit}\bar{Q}_n^*(\epsilon)(W) = \text{logit}\bar{Q}_n^0(W) + \epsilon H_{g_n}^*(W),$$

the clever covariate is defined as $H_{g_n}^*(W) = \frac{1}{g_n(1|W)}$, and ϵ, the fluctuation parameter, is estimated by maximum likelihood in which the loss function is thus the log-likelihood loss function:

$$-L(\bar{Q})(O) = \Delta\left\{Y \log \bar{Q}(W) + (1-Y)\log(1-\bar{Q}(W))\right\}. \tag{3.3}$$

Thus ϵ_n is fitted with univariate logistic regression, using the initial regression estimator $\bar{Q}_n^0$ as an off-set:

$$\epsilon_n = \arg\min_{\epsilon} \sum_{i=1}^{n} L(\bar{Q}_n^0(\epsilon))(O_i).$$

For estimators $\bar{Q}_n^0$ and g_n, one may specify a parametric model or use machine learning or even super learner, which uses loss-based cross-validation to select weighted combination of candidate estimators (van der Laan et al. [2007]).

Next, consider that Y is continuous, but bounded by 0 and 1. In this case, we can implement the same TMLE as we would for binary Y in (3.2). That is, we use the same logistic regression submodel, and the same loss function (3.3), and the same standard software

for logistic regression to fit ϵ, simply ignoring that Y is not binary. The same loss function is still valid for the conditional mean $\bar{Q}_0$ (Wedderburn [1974], Gruber and van der Laan [2010a]):

$$\bar{Q}_0 = \arg\min_{\bar{Q}} E_0 L(\bar{Q}).$$

Finally, given our continuous $Y \in [a, b]$, we can define $Y^* = (Y - a)/(b - a)$ so that $Y^* \in [0, 1]$. Then, let $\mu^*(P_0) = E_0(E_0(Y^* \mid \Delta = 1, W))$. We now compute the above TMLE of $\mu^*(P_0)$, and we use the relation $\mu(P_0) = (b - a)\mu^*(P_0) + a$.

We note that the previously proposed TMLE [Scharfstein et al., 1999] discussed in the KS debate would use the squared error loss function $L(\bar{Q})(O) = (Y - \bar{Q}(W))^2$, and the univariate linear regression model $\bar{Q}_n^0(\epsilon) = \bar{Q}_n^0 + \epsilon H_{g_n}^*$; however in this case, due to the fact that large values of the clever covariate map into equally large values of the regression function, global bounds may be violated, resulting in the loss of robustness. In our simulations, we include this TMLE, defined by the squared error loss function and linear fluctuation, as well as the TMLE defined by the quasi-log-likelihood loss function and the logistic fluctuation.

We note that our TMLE for continuous outcomes, that uses a squared error loss and linear fluctuation function, uses the same clever covariate as introduced by Scharfstein et al. [1999]. However, as also discussed in an addendum to Rosenblum and van der Laan [2010b], the Scharfstein et al. [1999] it is a special type TMLE due to using a clever parametric regression as initial estimator, thereby removing the need for the TMLE-update, but also restricting the estimator to parametric regression models. Both of these TMLE's (squared error loss and linear fluctuation) suffer from the same sensitivity to lack of positivity.

Finally, a natural extension of all of the above TMLE's is to make a more sophisticated estimate of g_0. Therefore, estimator $\mu_{n,C-TMLE}$ is defined by (3.2) as well, but the algorithm for computing Q_n^* differs. For the C-TMLE, we generate a sequence of nested-logistic regression model fits of g_0, $g_{n,1}, \ldots, g_{n,K}$, and we create a corresponding sequence of candidate TMLE's $Q_{k,g_{n,k}}^*$, using $g_{n,k}$ in the targeted MLE step, $k = 1, \ldots, K$, such that the loss-function (e.g., log-likelihood) specific fit of $Q_{k,g_{n,k}}^*$ is increasing in k. Finally, we use loss-function specific cross-validation to select k. The precise algorithm is presented in Gruber and van der Laan [2010b] and the software is available, and posted on *www.stat.berkeley.edu/ laan*. As a result, the resulting estimator g_n used in the TMLE is aimed to only include covariates that are effective in removing bias w.r.t. the target parameter: the theoretical underpinnings in terms of collaborative double robustness of the efficient influence curve is presented in van der Laan and Gruber [2009b].

3.5 Simulation Studies

In this section, we compare the performance of TMLE's to the estimating equation-based DR estimators (WLS, A-IPTW and BHT) as well as PRC and OLS, in the context of positivity violations. The goal of the original simulation designed by KS was to highlight the

stability problems of DR estimators, and they did this effectively. We wish to demonstrate the performance of the TMLE's in these simulations. We replicate the original KS simulation, and we also modify it in two ways, in order to explore the relative performance of the estimators under different and even more challenging data generating distributions.

3.5.1 Kang and Schafer Simulation

Kang and Schafer [2007] consider n i.i.d. units of $O = (W, \Delta, \Delta Y) \sim P_0$, where W is a vector of 4 baseline covariates, and Δ is an indicator of whether the continuous outcome, Y, is observed. Kang and Schafer are interested in estimating the following parameter:

$$\mu(P_0) = E_0(Y) = E_0(E_0(Y|\Delta = 1, W))$$

Let $(Z_1, \ldots, Z_4)$ be independent normally distributed random variables with mean zero and variance 1. The covariates W we actually observe are generated as follows:

$$\begin{aligned} W_1 &= exp(Z_1/2) \\ W_2 &= Z_2/(1 + exp(Z_1)) + 10 \\ W_3 &= (Z_1 Z_3/25 + 0.6)^3 \\ W_4 &= (Z_2 + Z_4 + 20)^2. \end{aligned}$$

The outcome Y is generated as

$$Y = 210 + 27.4Z_1 + 13.7Z_2 + 13.7Z_3 + 13.7Z_4 + N(0, 1).$$

From this one can determine that the conditional mean $\bar{Q}_0(W)$ of Y, given W, which equals the same linear regression in $Z_1(W), \ldots, Z_4(W)$, where $Z_j(W)$, $j = 1, \ldots, 4$, are the unique solutions of the 4 equations above in terms of $W = (W_1, \ldots, W_4)$. Thus, if the data analyst would have been provided the functions $Z_j(W)$, then the true regression function is linear in these functions, but the data analyst is measuring the terms W_j instead.

The other complication of the data generating distribution is that Y is subject to missingness, and the true censoring mechanism, denoted by $g_0(1|W) \equiv P_0(\Delta = 1|W)$, is given by:

$$g_0(1|W) = \text{expit}(-Z_1(W) + 0.5Z_2(W) - 0.25Z_3(W) - 0.1Z_4(W)).$$

With this data generating mechanism, the average response rate is 0.50. Also, the true population mean is 210, while the mean among respondents is 200. These values indicate a small selection bias.

In these simulations, a linear main term model in the main terms $(W_1, \ldots, W_4)$ for either the outcome-regression or missingness mechanism is misspecified, while a linear main term model in the main terms $(Z_1(W), \ldots, Z_4(W))$ would be correctly specified.

Note that in the KS simulation, there are finite sample violations of the positivity assumption. Specifically, we find $g_0(\Delta = 1|W) \in [0.01, 0.98]$ and the estimated missingness probabilities $g_n(\Delta = 1|W)$ were observed to fall in the range $[4x10^{-6}, 0.97]$.

3.5.2 Modification 1 of Kang and Schafer Simulation

In the KS simulation, when $\bar{Q}_0$ or g_0 are misspecified the misspecifications are small. The selection bias is also small. Therefore, we modified the KS simulation in order to increase the degree of misspecification and to increase the selection bias. This creates a greater challenge for estimators and better highlights their relative performance.

As before, let Z_j be i.i.d. $N(0,1)$. The outcome Y is generated as $Y = 210 + 50Z_1 + 25Z_2 + 25Z_3 + 25Z_4 + N(0,1)$. The covariates actually observed by the data analyst are now given by the following functions of $(Z_1, \ldots, Z_4)$:

$$\begin{aligned} W_1 &= exp(Z_1^2/2) \\ W_2 &= 0.5Z_2/(1 + exp(Z_1^2)) + 3 \\ W_3 &= (Z_1^2 Z_3/25 + 0.6)^3 + 2 \\ W_4 &= (Z_2 + 0.6Z_4)^2 + 2. \end{aligned}$$

From this one can determine the true regression function $\bar{Q}_0(W) = E_0(E(Y \mid Z) \mid W)$. The missingness indicator is generated as follows:

$$g_0(1|W) = \text{expit}(-2Z_1 + Z_2 - 0.5Z_3 - 0.2Z_4).$$

A misspecified fit is now obtained by fitting a linear or logistic main term regression in $W_1, \ldots, W_4$, while a correct fit is obtained by providing the user with the terms $Z_1, \ldots, Z_4$, and fitting a linear or logistic main term regression in $Z_1, \ldots, Z_4$. With these modifications, the population mean is again 210, but the mean among respondents is 184.4. With these modifications, we have a higher degree of practical violation of the positivity assumption: $g_0(\Delta = 1|W) \in [1.1x10^{-5}, 0.99]$ while the estimated probabilities, $g_n(\Delta = 1|W)$, were observed to fall in the range $[2.2x10^{-16}, 0.87]$.

3.5.3 Modification 2 of Kang and Schafer Simulation

For this simulation, we made one additional change to Modification 1: we set the coefficient in front of Z_4 in the true regression of Y on Z equal to zero. Therefore, while Z_4 is still associated with missingness, it is not associated with the outcome, and is thus not a confounder. Given $(W_1, \ldots, W_3)$, W_4 is not associated with the outcome either, and therefore as misspecified regression model of $\bar{Q}_0(W)$ we use a main term regression in (W_1, W_2, W_3).

This modification to the KS simulation enables us to take the debate on the relative performance of DR estimators one step further, by addressing a second key challenge of the estimators - that they often include non-confounders in the censoring mechanism estimator. This unnecessary inclusion could unnecessarily introduce positivity violations. Moreover, this unnecessary inclusion can itself introduce substantial bias and inflated variance, sometimes referred to as Z-bias. If the relationships between the variables are linear, the inclusion on non-confounders in the censoring mechanism will always increase bias [Bhattacharya and

Vogt, 2007, Wooldridge, 2009]. In the non-parametric case, the direction of the bias is less straightforward, but increasing bias is a real possibility [Pearl, 2010]. While this problem is not presented in the Kang and Schafer paper nor the responses, it is highlighted in the literature, including Bhattacharya and Vogt [2007], Wooldridge [2009] and Pearl [2010].

As discussed earlier, the C-TMLE algorithm provides an innovative black-box approach for estimating the censoring mechanism, preferring covariates that are associated with the outcome and censoring, without "data-snooping". With this modification to the KS simulation, we can compare C-TMLE to the other estimators when not all covariates are true confounders.

3.6 Results

For the three simulations described above, the OLS, WLS, A-IPCW, BHT, PRC, TML and C-TML estimators were used to estimate $\mu(P_0)$ from 250 samples of size 1000. We include the TMLE and C-TMLE based on the squared error loss function and linear regression submodel, as well as the TMLE (TMLEY*) and the C-TMLE (CTMLEY*) based on the quasi-log-likelihood loss function and logistic regression submodel. We evaluated the performance of the estimators by their bias, variance and mean squared error (MSE).

We compared the estimators of $\mu(P_0)$ using different specifications of the estimators of $\bar{Q}_0$ and g_0. In the tables below, "Qcgc" indicates that the estimators of both were specified correctly; "Qcgm" indicates that the estimator of $\bar{Q}_0$ was correctly specified, but the estimator of g_0 was misspecified; and "Qmgc" indicates that the estimator of $\bar{Q}_0$ was misspecified, but the estimator of g_0 was correctly specified. (Note that the co-authored paper also includes results for "Qmgm", indicating that both estimators were misspecified, but as as mentioned above, since this is not a focus of the comparisons in this chapter, these results have been currently omitted.)

For all estimators, we compared results with no lower bound on $g_n(1 \mid W)$ with truncating $g_n(1 \mid W)$ at three different lower bounds defined by three percentiles (0.01, 0.025, 0.05) of $g_n(1 \mid W_i)$, $i = 1, \ldots, n$. We note that neither KS nor Robins et al. [2007] included bounding $g_n(1 \mid W)$ when applying their estimators. Although, not bounding $g_n(1, W)$ has the advantage that in any given application it is difficult to determine which bounds to use, the theory teaches us that the DR estimators can only be consistent if g_n is bounded from below, even if in truth g_0 is unbounded.

Tables 1 to 3 present the simulation results without any bounding of g_n. The tables show that in all three simulations, the TMLE and C-TMLE with a logistic fluctuation achieve comparable or better MSE than the other estimators. When $\bar{Q}_n$ is misspecified, TMLEY* performs well and C-TMLEY* stands out with a much lower MSE. The tables show the importance of implementing a logistic fluctuation, as without it, the TMLE has substantial bias and variability if $\bar{Q}_n$ or g_n are misspecified. We note that in the case of "Qcgc", the TMLE's (except for TMLE) suffer from greater bias than the other estimators. This is due

Table 3.1: Simulation results with no bounding of g_n, Kang and Schafer simulation, 250 samples of size 1000

	Qcgc			**Qcgm**			**Qmgc**		
	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE
OLS	-0.092	1.40	1.41	-0.092	1.40	1.41	-0.927	1.97	2.83
WLS	-0.092	1.40	1.41	-0.092	1.41	1.42	0.099	1.84	1.85
A-IPCW	-0.092	1.40	1.41	-0.101	1.45	1.46	0.036	2.52	2.52
BHT	-0.092	1.40	1.41	-0.092	1.41	1.42	0.014	2.34	2.34
PRC	-0.092	1.40	1.41	-0.116	1.44	1.45	0.558	3.61	3.92
TMLE	-0.091	1.40	1.41	-0.094	1.39	1.40	0.095	2.52	2.53
TMLEY*	-0.149	1.40	1.43	-0.141	1.41	1.43	-0.160	1.92	1.94
C-TMLE	-0.156	1.40	1.43	-0.168	1.40	1.43	-0.327	1.81	1.92
C-TMLEY*	-0.159	1.40	1.43	-0.163	1.40	1.43	0.021	1.62	1.62

Table 3.2: Simulation results with no bounding of g_n, Modification 1 to Kang and Schafer simulation, 250 samples of size 1000

	Qcgc			**Qcgm**			**Qmgc**		
	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE
OLS	-0.165	4.69	4.72	-1.65e-01	4.69e00	4.72e00	-35.562	16.6	1281.2
WLS	-0.164	4.71	4.73	-1.64e-01	4.70e00	4.72e00	-4.401	41.9	61.3
A-IPCW	-0.162	4.75	4.77	-1.65e-01	4.69e00	4.72e00	-1.827	193.7	197.1
BHT	-0.162	4.73	4.76	-1.70e-01	4.71e00	4.74e00	-3.036	64.6	73.8
PRC	-0.177	4.74	4.77	6.80e08	1.78e21	1.78e21	80.641	8650.7	15153.7
TMLE	-0.169	4.71	4.74	-2.21e08	1.21e19	1.22e19	42.069	2402.6	4172.4
TMLEY*	-0.385	4.77	4.92	-4.27e-01	4.73e00	4.91e00	-0.607	68.8	69.2
C-TMLE	-0.448	4.77	4.97	-3.98e-01	4.73e00	4.89e00	4.490	96.1	116.3
C-TMLEY*	-0.440	4.78	4.97	-4.13e-01	4.77e00	4.94e00	-0.956	10.8	11.8

to bounding the predicted values by the range of the observed outcome, which is skewed due to the informed missingness. One approach that can reduce this bias is to bound by a range slightly larger (i.e. 10 percent) than the true range of the observed $Y's$.

Together, the results from Modification 1 and Modification 2 show that the C-TMLE's have similar or superior performance relative to estimating equation-based DR estimators when not all covariates are associated with Y. At the same time, even in cases in which *all* covariates are associated with Y, C-TMLE's still perform well.

Tables 4 to 6 compare results for each estimator when bounding g_n at different levels. We observe that bounding g_n can improve the bias and variability of the estimators, often substantially. However, we also see that bounding can easily increase bias. The effect of bounding and the desired level of bounding varies by estimator.

It is more important to note that C-TMLEY* and TMLEY* are always well behaved.

Table 3.3: Simulation results with no bounding of g_n, Modification 2 to Kang and Schafer simulation, 250 samples of size 1000

	Qcgc			**Qcgm**			**Qmgc**		
	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE
OLS	-0.058	3.93	3.94	-5.75e-02	3.93e00	3.94e00	-34.260	15.26	1189.01
WLS	-0.057	3.96	3.96	-5.61e-02	3.94e00	3.94e00	-4.021	37.75	53.91
A-IPCW	-0.0545	3.99	4.00	-5.68e-02	3.94e00	3.94e00	-1.248	172.94	174.49
BHT	-0.055	3.98	3.98	-5.86e-02	3.96e00	3.97e00	-2.406	67.60	73.39
PRC	-0.069	4.04	4.04	6.80e08	1.78e21	1.78e21	78.135	7772.57	13877.69
TMLE	-0.061	3.97	3.98	-2.21e08	1.21e19	1.22e19	40.920	2136.03	3810.49
TMLEY*	-0.274	3.97	4.05	-3.08e-01	3.91e00	4.01e00	-0.451	59.04	59.25
C-TMLE	-0.331	3.98	4.09	-2.77e-01	3.94e00	4.02e00	4.356	76.01	94.99
C-TMLEY*	-0.321	3.98	4.08	-2.97e-01	3.93e00	4.02e00	-1.175	7.49	8.87

In no simulation do they show marked instability. C-TMLEY* performs particularly well.

Table 3.4: Simulation results, bounding g_n, KS simulation, 250 samples of size 1000

Lower	**Qcgc**			**Qcgm**			**Qmgc**		
g_n	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE
OLS									
None	-0.0921	1.40	1.41	-0.0921	1.40	1.41	-0.9266	1.97	2.83
0.01	-0.0921	1.40	1.41	-0.0921	1.40	1.41	-0.9266	1.97	2.83
0.05	-0.0921	1.40	1.41	-0.0921	1.40	1.41	-0.9266	1.97	2.83
WLS									
None	-0.0915	1.40	1.41	-0.0916	1.41	1.42	0.0986	1.84	1.85
0.01	-0.0915	1.40	1.41	-0.0913	1.41	1.42	0.0986	1.84	1.85
0.05	-0.0915	1.40	1.41	-0.0920	1.41	1.41	0.1118	1.81	1.83
A-IPCW									
None	-0.0915	1.40	1.41	-0.1014	1.45	1.46	0.0361	2.52	2.52
0.01	-0.0915	1.40	1.41	-0.0913	1.41	1.42	0.0361	2.52	2.52
0.05	-0.0914	1.40	1.41	-0.0922	1.41	1.42	0.0807	2.28	2.29
BHT									
None	-0.0915	1.40	1.41	-0.0918	1.41	1.42	0.0139	2.34	2.34
0.01	-0.0915	1.40	1.41	-0.0916	1.41	1.42	0.0139	2.34	2.34
0.05	-0.0914	1.40	1.41	-0.0917	1.41	1.41	0.0877	1.98	1.99
PRC									
None	-0.0920	1.40	1.41	-0.1162	1.44	1.45	0.5584	3.61	3.92
0.01	-0.0921	1.40	1.41	-0.0940	1.40	1.41	0.5554	3.60	3.91
0.05	-0.0921	1.40	1.41	-0.0926	1.40	1.41	0.3939	3.01	3.16
TMLE									
None	-0.0914	1.40	1.41	-0.0944	1.39	1.40	0.0952	2.52	2.53
0.01	-0.0914	1.40	1.41	-0.0926	1.41	1.42	0.0948	2.52	2.52
0.05	-0.0915	1.40	1.41	-0.0923	1.41	1.41	0.1161	2.34	2.36
TMLEY*									
None	-0.1494	1.40	1.43	-0.1412	1.41	1.43	-0.1603	1.92	1.94
0.01	-0.1495	1.40	1.43	-0.1426	1.41	1.43	-0.1603	1.92	1.94
0.05	-0.1500	1.40	1.43	-0.1482	1.41	1.43	-0.1461	1.86	1.88
C-TMLE									
None	-0.1556	1.40	1.43	-0.1678	1.40	1.43	-0.3271	1.81	1.92
0.01	-0.1548	1.40	1.43	-0.1561	1.41	1.43	-0.3245	1.81	1.91
0.05	-0.1544	1.40	1.43	-0.1556	1.40	1.43	-0.2819	1.81	1.89
C-TMLEY*									
None	-0.1589	1.40	1.43	-0.1633	1.40	1.46	0.0213	1.62	1.62
0.01	-0.1587	1.41	1.43	-0.1604	1.41	1.46	0.0208	1.65	1.65
0.05	-0.1586	1.40	1.43	-0.1591	1.41	1.47	0.0162	1.64	1.64

Table 3.5: Simulation results, bounding g_n, Modification 1 to KS simulation, 250 samples of size 1000

Lower	**Qcgc**			**Qcgm**			**Qmgc**		
g_n	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE
OLS									
None	-0.165	4.69	4.72	-1.65e-01	4.69e00	4.72e00	-35.562	16.58	1281.2
0.01	-0.165	4.69	4.72	-1.65e-01	4.69e00	4.72e00	-35.562	16.58	1281.2
0.05	-0.165	4.69	4.72	-1.65e-01	4.69e00	4.72e00	-35.562	16.58	1281.2
WLS									
None	-0.164	4.71	4.73	-1.64e-01	4.70e00	4.72e00	-4.401	41.95	61.3
0.01	-0.165	4.71	4.73	-1.64e-01	4.70e00	4.72e00	-4.609	38.93	60.2
0.05	-0.166	4.70	4.73	-1.64e-01	4.70e00	4.72e00	-7.341	24.46	78.4
A-IPCW									
None	-0.162	4.75	4.77	-1.65e-01	4.69e00	4.72e00	-1.827	193.73	197.1
0.01	-0.165	4.71	4.74	-1.65e-01	4.69e00	4.72e00	-3.665	74.34	87.8
0.05	-0.166	4.70	4.73	-1.65e-01	4.69e00	4.72e00	-8.817	27.85	105.6
BHT									
None	-0.162	4.73	4.76	-1.70e-01	4.71e00	4.74e00	-3.036	64.63	73.8
0.01	-0.164	4.71	4.74	-1.69e-01	4.71e00	4.74e00	-3.579	49.69	62.5
0.05	-0.166	4.70	4.73	-1.66e-01	4.71e00	4.74e00	-7.179	23.62	75.2
PRC									
None	-0.177	4.74	4.77	6.80e08	1.78e21	1.78e21	80.641	8650.67	15153.7
0.01	-0.164	4.70	4.73	-1.65e-01	4.70e00	4.72e00	20.820	870.27	1303.7
0.05	-0.164	4.70	4.72	-1.65e-01	4.70e00	4.72e00	6.301	61.46	101.2
TMLE									
None	-0.169	4.71	4.74	-2.21e08	1.21e19	1.22e19	42.069	2402.58	4172.4
0.01	-0.164	4.71	4.73	-1.65e-01	4.69e00	4.72e00	7.691	210.00	269.1
0.05	-0.165	4.70	4.73	-1.65e-01	4.69e00	4.72e00	-0.951	22.80	23.7
TMLEY*									
None	-0.385	4.77	4.92	-4.27e-01	4.73e00	4.91e00	-0.607	68.78	69.2
0.01	-0.392	4.78	4.93	-4.16e-01	4.77e00	4.94e00	0.135	39.63	39.6
0.05	-0.402	4.77	4.94	-4.15e-01	4.77e00	4.94e00	-0.769	10.24	10.8
C-TMLE									
None	-0.448	4.77	4.97	-3.98e-01	4.73e00	4.89e00	4.490	96.10	116.3
0.01	-0.420	4.77	4.95	-3.95e-01	4.76e00	4.91e00	3.956	64.69	80.3
0.05	-0.411	4.76	4.93	-3.95e-01	4.76e00	4.91e00	-1.839	9.46	12.8
C-TMLEY*									
None	-0.440	4.78	4.97	-4.13e-01	4.77e00	4.94e00	-0.956	10.84	11.8
0.01	-0.429	4.77	4.96	-4.13e-01	4.77e00	4.94e00	-1.031	11.04	12.1
0.05	-0.422	4.77	4.95	-4.13e-01	4.77e00	4.94e00	-2.535	8.49	14.9

Table 3.6: Simulation results, bounding g_n, Modification 2 to KS simulation, 250 samples of size 1000

Lower	**Qcgc**			**Qcgm**			**Qmgc**		
g_n	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE
OLS									
None	-0.058	3.93	3.94	-5.75e-02	3.93e00	3.94e00	-34.260	15.26	1189.01
0.01	-0.058	3.93	3.94	-5.75e-02	3.93e00	3.94e00	-34.260	15.26	1189.01
0.05	-0.058	3.93	3.94	-5.75e-02	3.93e00	3.94e00	-34.260	15.26	1189.01
WLS									
None	-0.057	3.96	3.96	-5.61e-02	3.94e00	3.94e00	-4.021	37.75	53.91
0.01	-0.057	3.96	3.96	-5.61e-02	3.94e00	3.94e00	-4.238	34.54	52.49
0.05	-0.058	3.95	3.96	-5.62e-02	3.94e00	3.94e00	-6.876	22.49	69.77
A-IPCW									
None	-0.055	3.99	4.00	-5.68e-02	3.94e00	3.94e00	-1.248	172.94	174.49
0.01	-0.057	3.96	3.97	-5.68e-02	3.94e00	3.94e00	-3.190	69.18	79.36
0.05	-0.058	3.95	3.95	-5.69e-02	3.94e00	3.94e00	-8.348	25.61	95.29
BHT									
None	-0.055	3.98	3.98	-5.86e-02	3.96e00	3.97e00	-2.406	67.60	73.39
0.01	-0.057	3.96	3.96	-5.84e-02	3.98e00	3.98e00	-3.132	45.53	55.34
0.05	-0.058	3.95	3.95	-5.74e-02	3.96e00	3.96e00	-6.768	21.57	67.37
PRC									
None	-0.069	4.04	4.04	6.80e08	1.78e21	1.78e21	78.135	7772.57	13877.69
0.01	-0.056	3.96	3.97	-5.68e-02	3.94e00	3.94e00	20.389	798.43	1214.13
0.05	-0.057	3.95	3.95	-5.70e-02	3.94e00	3.94e00	6.274	56.28	95.64
TMLE									
None	-0.061	3.97	3.98	-2.21e08	1.21e19	1.22e19	40.920	2136.03	3810.49
0.01	-0.057	3.96	3.96	-5.68e-02	3.94e00	3.94e00	7.714	189.96	249.47
0.05	-0.058	3.95	3.95	-5.69e-02	3.94e00	3.94e00	-0.738	20.58	21.12
TMLEY*									
None	-0.274	3.97	4.05	-3.08e-01	3.91e00	4.01e00	-0.451	59.04	59.25
0.01	-0.278	3.97	4.05	-2.96e-01	3.95e00	4.04e00	0.353	31.18	31.31
0.05	-0.285	3.97	4.05	-2.96e-01	3.95e00	4.04e00	-0.554	8.15	8.46
C-TMLE									
None	-0.331	3.98	4.09	-2.77e-01	3.94e00	4.02e00	4.356	76.01	94.99
0.01	-0.303	3.96	4.06	-2.76e-01	3.95e00	4.02e00	3.660	53.25	66.65
0.05	-0.293	3.96	4.05	-2.76e-01	3.95e00	4.02e00	-1.708	7.40	10.32
C-TMLEY*									
None	-0.3213	3.98	4.08	-2.97e-01	3.93e00	4.02e00	-1.175	7.49	8.87
0.01	-0.3115	3.97	4.07	-2.97e-01	3.93e00	4.02e00	-1.257	6.71	8.29
0.05	-0.3044	3.96	4.05	-2.93e-01	3.93e00	4.04e00	-2.459	5.90	11.95

3.7 TMLE's with Machine Learning for Dual Misspecification

In this section, we couple super learning with TMLE and C-TMLE to estimate both $\bar{Q}_0$ and g_0. For C-TMLEY*, four missingness-mechanism score-based covariates were created based on different truncation levels of the propensity score estimate $g_n(1 \mid W)$: no truncation, and truncation from below at the 0.01, 0.025, and 0.05-percentile. These four scores were supplied along with the misspecified main terms $W_1, \ldots, W_4$ to the targeted forward selection algorithm in the C-TMLEY* used to build a series of candidate nested logistic regression estimators of the missingness mechanism and corresponding candidate TMLE's. The C-TMLEY* algorithm used 5-fold cross-validation to select the best estimate from the eight candidate TMLE's. This allows the C-TMLE algorithm to build a logistic regression fit of g_0 that selects among the misspecified main-terms and super-learning fits of the missingness mechanism score $g_n(1 \mid W)$ at different truncation levels.

An important aspect of super learning is to ensure that the library of prediction algorithms includes a variety of approaches for fitting the true function $\bar{Q}_0$ and g_0. For example, it is sensible to include a main terms regression algorithm in the super learner library. Should that algorithm happen to be correct, the super learner will behave as the main terms regression algorithm. It is also recommended to include algorithms that search over a space of higher order polynomials, non-linear models, and, for example, cubic splines. For binary outcome regression, as required for fitting g_0, classification algorithms such as classification and regression trees [Breiman et al., 1984], support vector machines [Cortes and Vapnik, 1995]), and k-nearest-neighbor algorithms (Friedman [1994]), could be added to the library. The point of super-learning is that we cannot know in advance which procedure will be most successful for a given prediction problem. Super learning relies on the oracle property of V-fold cross-validation to asymptotically select the optimal convex combination of estimates obtained from these disparate procedures (van der Laan and Dudoit [2003a], van der Laan et al. [2004a], van der Laan et al. [2007]).

Consider the misspecified scenario proposed by KS. The true full-data distribution and the missingness mechanism are captured by main terms linear regression of the outcome on Z_1, Z_2, Z_3, Z_4. This simple model is virtually impossible to discover through the usual model selection approaches when the observed data consists of misspecified covariates $O = (W_1, W_2, W_3, W_4, \Delta, \Delta Y)$, given that

$$\begin{aligned} Z_1 &= 2log(W_1), \\ Z_2 &= (W_2 - 10)(1 + 2W_1), \\ Z_3 &= \frac{25(W_3 - 0.6)}{2log(W_1)}, \\ Z_4 &= \sqrt[3]{W_4} - 20 - (W_2 - 10)(1 + 2W_1). \end{aligned}$$

This complexity illustrates the importance of including prediction algorithms that attack

the estimation problem from a variety of directions. The super learner library we employed contained the algorithms listed below. The analysis was carried out in the R statistical programming environment v2.10.1 [Team, 2010], using algorithms included in the base installation or in the indicated package.

- **glm** (base) main terms linear regression.
- **step** (base) stepwise forward and backward selection using the AIC criterion [Hastie and Pregibon, 1992].
- **ipredbagg** (ipred) bagging for classification, regression and survival trees [Peters and Hothorn, 2009, Breiman, 1996].
- **DSA** (DSA) Deletion/Selection/Addition algorithm for searching over a space of polynomial models or order k (k set to 2). [Neugebauer and Bullard, 2010, Sinisi and van der Laan, 2004]
- **earth** (earth) Building a regression model using multivariate adaptive regression splines (MARS) [Milborrow, 2009, Friedman, 1991, 1993].
- **loess** (stats) Local polynomial regression fitting [W.S. Cleveland and Shyu, 1992].
- **nnet** (nnet) Single-hidden-layer neural network for classification [Ripley, 1996].
- **svm** (e1071) Support vector machine for regression and classification [Dimitriadou et al., 2010, Chang and Lin, 2001].
- **k-nearest-neighbors*** (class) classification using most common outcome among identified k nearest nodes (k set to 10) [Venables and Ripley, 2002, Friedman, 1994]

* only for binary outcomes, added to library for estimating g

3.7.1 Results

In table 3.7 we report the results for TMLEY* and CTMLEY* based on 250 samples of size 1000, with predicted values for $g_n(1 \mid W)$ truncated from below at 0.025. The MSE for both estimators is smaller than the MSE of $\hat{\mu}_{OLS}$. C-TMLEY^* bias is slightly higher than $\hat{\mu}_O LS$ bias, and TMLEY^* is slightly better with respect to both bias and variance. More importantly, the data-adaptive estimation approach improved efficiency of TMLEY* by a factor of 8.5. C-TMLEY* efficiency improved by a factor of 1.5.

Table 3.7: Results incorporating super learning into TMLE and C-TMLE, with $g_n(1 \mid W)$ truncated at 0.025

	Bias	**Var**	**MSE**
TMLEY* + SL	-0.771	1.51	2.10
CTMLEY* + SL	-1.047	1.54	2.64

3.8 Discussion

By mapping continuous outcomes into [0,1] and using a logistic fluctuation, we show that the TMLE's (both TMLEY* and C-TMLEY*) are more robust to violations of the positivity assumption than the TMLE's using the linear fluctuation function. We also show that C-TMLE's have superior performance relative to estimating equation-based DR estimators when not all covariates are associated with the outcome Y. The C-TMLE algorithm provides an innovative approach for estimating the censoring mechanism, preferring covariates that are associated with the outcome Y and missingness, Δ. C-TMLE's avoid data snooping concerns because the estimation procedure is fully specified before the analyst observes any data (or at least, not any data beyond some ancillary statistics). Even in cases in which *all* observed covariates are associated with Y, C-TMLE's still perform well.

Other recent work has also investigated the relative performance of DR estimators using the KS simulations. Cao et al. [2009] and Tan [2010] use the simulation to compare numerous existing non-DR and DR estimators to alternative estimators that they introduce. Cao et al. [2009] present an estimator that achieves minimum variance when the estimator of missingness mechanism is correctly specified (see also Rubin and van der Laan [2008] for empirical efficiency maximization), and they address the effect of large IPCW by enhancing the missingness mechanism estimator in order to constrain the predicted values. They demonstrate that their estimators perform comparably to existing estimators when $\bar{Q}_n$ is correct, but outperform the others when $\bar{Q}_n$ is misspecified and g_n is correct, as well as when both are misspecified. No TMLE's are included in their comparisons.

Tan [2010] presents the "calibrated likelihood estimator" and the "augmented likelihood estimator," both more robust versions of estimators originally introduced in Tan [2006]. The first of these two estimators respects global bounds and is semi-parametric efficient, while the second respects a weaker form of boundedness. Tan [2010] finds that these two estimators achieve the lowest MSE in the simulation when compared to numerous existing estimators. Included in the comparisons is the TMLE for a continuous Y using the linear fluctuation function. Tan [2010] does not include the estimators in Cao et al. [2009] in the comparisons.

Related work is also being done with respect to other parameters of interest. Both Cao et al. [2009] and Tan [2006] include discussions on applying their estimators to causal effect parameters. In addition, Freedman and Berk [2008], focus on a causal effect parameter,

and demonstrate that DR estimators (and the WLS estimator in particular) can increase variance and bias when IPCW are large.

Overall, comparisons of estimators, beyond theoretical studies of asymptotics as well as robustness, will need to be based on large scale simulation studies, including all available estimators, and cannot be tailored towards one particular simulation setting. Future research should be concerned with setting up such a large scale objective comparison based on publicly available software, and we are looking forward to contribute to such an effort.

The research underlying TMLE's was motivated, in part, by the goal of increasing the stability of DR estimators, and the KS simulations provide a demonstration of the merits of TMLE's under violations of the positivity assumption. TMLE's are estimators defined by the choice of loss function, and parametric submodel, both chosen so that the linear span of the scores at zero fluctuation w.r.t. the loss function includes the efficient influence curve/efficient score. All such TMLE's are double robust, asymptotically efficient under correct specification, and substitution estimators, but the choice of loss function and submodel can affect the finite sample robustness, as observed in the current simulations. In addition, TMLE's can be combined with super learning and empirical efficiency maximization (Rubin and van der Laan [2008] and van der Laan and Gruber [2009b]) to further enhance their performance in practice. We hope that by showing that these estimators perform well in simulations and settings created by *other* researchers, for the purposes of showing the weaknesses of DR estimators, as well as in modified simulations that make estimation even more challenging, we provide probative evidence in support of TMLE's. Of course, much can happen in finite samples, and we look forward to further exploring how these estimators perform in other settings.

Chapter 4

Targeted Maximum Likelihood Estimation of Conditional Relative Risk Parameters in a Semi-parametric Multiplicative Regression Model

4.1 Introduction

The first two chapters of this thesis illustrated the challenges of estimating two different parameters under violations of the positivity assumption and demonstrated the improved robustness one can obtain by using TMLE's, particularly when incorporating a logistic fluctuation. The first chapter also included a discussion of approaches for responding to bias due to positivity violations for any given estimator. The approaches all involved making trade-offs between proximity to the initial target parameter and identifiability. Another approach for responding to positivity bias that was not mentioned earlier is to introduce model assumptions, thereby modifying the statistical model. This does not necessarily change the parameter of interest, as the approaches in Chapter 1 do, except that if the model we assume is not correct, we are not estimating the parameter we intended. For this reason, making model assumptions is not ideal. However, we can use semi-parametric methods to maximize flexibility. In addition, we can implement targeted maximum likelihood estimation for parameters in the semi-parametric model. As a result, assuming the semi-parametric model results in a correct model, we can obtain robust estimators under positivity violations.

In this chapter, we demonstrate the value of this approach, focusing on a new parameter compared to those presented in Chapters 1 and 2. In particular, we introduce a semi-parametric multiplicative regression model for a binary outcome. We then consider two TMLE's of the parameter of interest in the assumed model (which may be a vector), a parameter that relates exposure to changes in conditional relative. One of these two TMLE's

correctly assumes that the binary outcome (e.g. disease or no disease) has a binomial distribution. This results in a double-robust (DR), efficient estimator of the parameter of interest in the model, but it is unstable, due to convergence problems with the log-binomial regression model, which is used for estimation. The second TMLE instead assumes that the outcome is a count of events and follows a Poisson distribution. If the outcome truly did follow a Poisson distribution, this TMLE would also be a DR, efficient estimator of the parameter of interest, which would then relate exposure to changes in the conditional incidence rate rather than to changes in conditional relative risk. However, we apply the second TMLE to data in which the outcome is binary and therefore the assumption of a Poisson distribution is always wrong. In this case, the TMLE is no longer efficient, but it does achieve stability. It also remains DR - that is, the efficient score estimating function in the semi-parametric Poisson regression model is an unbiased DR estimating function for the parameter of interest in the semi-parametric conditional mean model, which does not assume a Poisson distribution. Consequently, this second TMLE is consistent and we can provide correct inference. We refer to this latter TMLE as the "practical TMLE" of our parameter of interest when the outcome in truly binary, and we focus on this TMLE in our implementation and simulations in this paper.

The layout of this chapter is as follows. In Section 4.2, we present the data structure. In Section 4.3, we present the semi-parametric multiplicative model and formalize the parameter of interest, which is implied by the model. Then in Section 4.4, we introduce the two TMLE's of the parameter in the semi-parametric multiplicative regression model - the "correct" TMLE and the "practical" TMLE. In this section we also describe how the estimation procedure for both TMLE's can be easily modified for case-control data. Section 4.5 provides a step-by-step implementation for the "practical" TMLE in either a prospective or case-control sample, and in Section 4.6, we demonstrate the performance of this TMLE with results from a variety of simulations. This section includes comparisons to common estimation methods in the epidemiology and medical literature. Finally, a brief summary discussion is provided in Section 4.7. The final chapter of this thesis provides methodological details for the TMLE's discussed in this paper. We note that the content of this chapter is summarized in Tuglus et al. [2011].

4.2 Data Structure

Consider an observed point treatment data set consisting of n independent and identically distributed (i.i.d.) observations of $O = (W, A, Y) \sim P_0 \in \mathcal{M}$. W is a vector of baseline covariates, A is the exposure of interest and $Y = \{0, 1\}$ is a binary outcome. P_0 denotes the true distribution of O, from which all subjects are sampled. P_0 is an element of a statistical model $\mathcal{M}$, which is a semi-parametric model defined below in Section 4.3.

Note that for causal effects, we assume that O is a missing data structure on a hypothetical full data structure $X = (W, Y_a : a \in A)$, which contains all counterfactual outcomes

Y_a. We therefore view A as the missingness variable, as O contains only one of all possible counterfactual outcomes, $Y = Y_A$.

4.3 Semi-parametric Multiplicative Model and Parameter of Interest

We define the statistical model with the following semi-parametric multiplicative model:

$$\bar{Q}_0(A, W) = e^{m_{\beta_0}(A,V)}\theta_0(W), \tag{4.1}$$

where $\bar{Q}_0(A, W) \equiv P_0(Y = 1|A, W)$, $m_{\beta_0}(A, V)$ is a specified function of A and effect modifiers $V \subset W$, and $\theta_0(W) \equiv P_0(Y = 1|A = 0, W)$ is a non-parametric model of the conditional expectation of Y given no exposure and baseline covariates W. For $m_{\beta_0}(A, V)$, we generally specify a linear function of A such as (1) $m_{\beta_0}(A, V) = \beta_0 A$ or (2) $m_{\beta_0}(A, V) = \beta_{0(1)}A + \beta_{0(2)}A : V$, where the importance of the exposure is modified by covariate V. In this chapter, we primarily focus on the former, where $m_{\beta_0}(A, V) = \beta_0 A$. In this case, we can write

$$e^{\beta_0 A} = \frac{\bar{Q}_0(A, W)}{\theta_0(W)}, \tag{4.2}$$

or

$$\beta_0 A = \log\left\{\frac{\bar{Q}_0(A, W)}{\theta_0(W)}\right\}.$$

The parameter of interest, $\beta_0 = \Psi(\bar{Q}_0)$, which is implied by the model in (4.1) is then equivalent to the change in the log conditional relative risk associated with a unit increase of exposure A.

One motivation for this semi-parametric multiplicative model is that regardless of whether or not we believe our model, we can still accurately test the following strong null hypothesis:

$$H_0 : \frac{P_0(Y = 1|A, W)}{P_0(Y = 1|A = 0, W)} = 1,$$

for all W. Under this null, the model is always correct. Therefore, we can construct valid hypotheses tests.

We note that the model, as seen in (4.2), requires that both $\bar{Q}_0(A, W) > 0$ and $\theta_0(W) > 0$. In order for this to be true, we must have that $P_0(A = a|W) > 0$ for all a, w. This latter requirement is often referred to as the positivity assumption (Robins [1986, 1987a, 1999]), or the experimental treatment assignment (ETA) assumption (Neugebauer and van der Laan [2005]). Therefore, (4.2) is true only for a, w for which there is support in the data. So returning to the general case for any $m_{\beta_0}(A, V)$ and to be more precise, we can write

$$e^{m_{\beta_0}(a,v)} = \frac{\bar{Q}_0(a,w)}{\theta_0(w)} I(a,w \in \mathcal{A}'),$$

where $\mathcal{A}'$ contains the subset of a, w for which the positivity assumption is not violated (i.e. for which $P_0(A = a|W) > 0$). This allows us to estimate the importance (i.e. risk) of exposure A in predicting the outcome Y, conditional on W, for those strata of W in which the data have sufficient support. The analyst may want to allow for extrapolation across *all* a, w, but this would be unwise if the model was not true for all a, w.

In order for the parameter of interest to have a causal interpretation, we require that not only that O is a missing data structure on a hypothetical full data structure $X = (W, Y_0, Y_1)$, as described in Section 4.2, but we also require the randomization assumption: $\{A \perp Y_0, Y_1|W\}$. However, even in the case where these assumptions do not hold, the parameter β_0 is still a well-defined and meaningful parameter of variable importance (i.e. association adjusted for W.)

4.4 Targeted Maximum Likelihood Estimation

In this section, we introduce two TMLE's of β_0 in the given semi-parametric multiplicative regression model in (4.1). As mentioned above, the first TMLE correctly assumes that the conditional distribution of Y follows a binomial distribution and the second which incorrectly assumes it follows a Poisson distribution. Like the TMLE's introduced in earlier chapters, and all TMLE's, the TMLE's of β_0 in our semi-parametric multiplicative model are double robust (DR) and asymptotically linear, substitution estimators, which involve two steps: (1) obtaining an initial estimator of the likelihood of the data (or in the current case, an estimator of $\bar{Q}_0 \equiv P_0(Y = 1|A, W)$), and (2) updating this initial estimator in order to target bias reduction in the parameter of interest. Below, we describe each of these steps and how they differ for each of the TMLE's.

4.4.1 Initial Estimator, $\bar{Q}_n^0$

To obtain the initial estimator of $\bar{Q}_0$, we first need an estimator of θ_0. We do this by using a pre-specified data-adaptive estimator, preferably super learner, to obtain an estimator of $P_0(Y = 1|A, W)$ of general model form. As noted in earlier chapters, super learner takes a comprehensive library of data adaptive estimators and uses cross-validation to combine these estimators into at least an equal, but more often a better, estimator than any individual candidate in the library (van der Laan et al. [2007]). Using the resulting fit, we obtain θ_n by predicting the outcome given the observed data with all observations set to $A = 0$. We then fit the model $e^{m_{\beta_0}(A,W)}\theta_0(W)$ using a parametric regression model (i.e. log-binomial or Poisson). This gives us an initial estimator, $\bar{Q}_n^0(A, W)$ of the correct semi-parametric model form, with its corresponding estimate β_n^0 by substitution.

4.4.2 Updated, Targeted Estimator, $\bar{Q}_n^*$

Once we have our initial estimator $\bar{Q}_n^0$, we want to update it in order to target bias reduction in β_n^0 due to residual confounding and non-targeting. In other words, we want to find $\bar{Q}_n^*$ such that the TMLE $\beta_n^* = \Psi(\bar{Q}_n^*)$ is an unbiased estimator of β_0. We find $\bar{Q}_n^*$ by fluctuating the initial estimators of $\bar{Q}_0$ and θ_0. Therefore, we fluctuate our initial estimator, $\bar{Q}_n^0$ by defining a parametric submodel of our original model, with fluctuation parameter ϵ. The general parametric fluctuation submodel (without yet making any particular parametric assumptions about the distribution of Y) is given by

$$\log\bar{Q}_{\beta_n^0,\theta_n^0}(\epsilon)(A,W) = (\beta_n^0 + \epsilon)A + \log\theta_n^0(\epsilon)(W).$$

where $\theta_n^0(\epsilon)(W) = \theta_n^0(W)\exp(\epsilon r^*_{\bar{Q}_n^0,g_n}(W))$. The function $r^*_{\bar{Q}_n^0,g_n}(W)$ depends on $\bar{Q}_n^0$ (i.e. on β_n^0 and θ_n^0), as well as on $g_n(A|W)$, an estimator of the nuisance parameter $g_0(A|W) \equiv P_0(A|W)$. Therefore, we have

$$\begin{aligned} \log\bar{Q}_{\beta_n^0,\theta_n^0}(\epsilon)(A,W) &= (\beta_n^0 + \epsilon)A + \log\theta_n^0(W) + \epsilon r^*_{\bar{Q}_n^0,g_n}(W) \\ &= \beta_n^0 A + \log\theta_n^0(W) + \epsilon(A + r^*_{\bar{Q}_n^0,g_n}(W)). \end{aligned} \tag{4.3}$$

Therefore, the fluctuation parameter ϵ is a coefficient on an "clever covariate" given by $A + r^*_{\bar{Q}_n^0,g_n}(W)$, which is added to the log-linear model. To construct $r^*_{\bar{Q}_n^0,g_n}(W)$, we require that when $\epsilon = 0$, we have the initial density, $\bar{Q}_n^0(A,W)$ and we have that the score of the submodel corresponding with this choice $r^*_{\bar{Q}_n^0,g_n}(W)$ is equal to the efficient score of the parameter of interest, β_0. The efficient score and therefore the clever covariate will depend on our assumptions about the distribution of $\bar{Q}_0$. For this reason, depending on whether we assume a binomial or Poisson density, we derive a different TMLE. We provide more details for each TMLE below.

For both TMLE's, we fit ϵ with maximum likelihood, which results in a first step update of $\bar{Q}_n^0$, and we iteratively update an initial estimator until the next MLE of ϵ is close to zero. In other words, for each of $k = 0 \ldots K$ iterations, we compute an updated $\bar{Q}_{\beta_n^k,\theta_n^k}(\epsilon_n^k)$, using the function $r^*_{\bar{Q}_n^k,g_n}$. At convergence, the final updated estimator of $\bar{Q}_0$ (for either outcome type/parameter) is given by $\bar{Q}_n^* \equiv \bar{Q}_n^K$ and we evaluate β_n^* by substitution.

TMLE 1: Correctly assuming Y is binary and has a binomial distribution

When we correctly assume that $P_0(Y|A,W)$ follows a binomial distribution, the efficient score for the parameter of interest, β_0, in our semi-parametric multiplicative model is given by the following when we let $m_{\beta_0}(A,V) = \beta_0 A$:

$$S^*_{\bar{Q}_0,g_0}(O) = \frac{1}{1-\bar{Q}_0}\left(A - \frac{E_0\left[\frac{A\bar{Q}_0}{(1-\bar{Q}_0)}|W\right]}{E_0\left[\frac{\bar{Q}_0}{1-\bar{Q}_0}|W\right]}\right)(Y-\bar{Q}_0).$$

Then to find $r^*_{\bar{Q}_0,g_0}$, we arrange that the score of the fluctuation model at $\epsilon = 0$ equals this efficient score. We can then compute the clever covariate, which is given by

$$H^*_{\bar{Q}_0,g_0} = A - \frac{E_0\left[\frac{A\bar{Q}_0}{(1-\bar{Q}_0)}|W\right]}{E_0\left[\frac{\bar{Q}_0}{1-\bar{Q}_0}|W\right]}.$$

Detailed derivations for the above efficient score, $r^*_{\bar{Q}_0,g_0}$ and $H^*_{\bar{Q}_0,g_0}$ are provided in the next chapter. For the more general case, the efficient score is given by

$$S^*_{\bar{Q}_0,g_0}(O) = \frac{1}{1-\bar{Q}_0}\left(\frac{d}{d\beta_0}m_{\beta_0} - \frac{E_0\left[\frac{\frac{d}{d\beta_0}m_{\beta_0}\bar{Q}_0}{(1-\bar{Q}_0)}|W\right]}{E_0\left[\frac{\bar{Q}_0}{1-\bar{Q}_0}|W\right]}\right)(Y-\bar{Q}_0).$$

We note that this can also be represented as

$$S^*_{\bar{Q}_0,g_0}(O) = h^*(A \mid W)(Y-\bar{Q}_0),$$

where

$$h^*(A \mid W) = \frac{1}{1-\bar{Q}_0}\left(\frac{d}{d\beta_0}m_{\beta_0} - \frac{E_0\left[\frac{\frac{d}{d\beta_0}m_{\beta_0}\bar{Q}_0}{(1-\bar{Q}_0)}|W\right]}{E_0\left[\frac{\bar{Q}_0}{1-\bar{Q}_0}|W\right]}\right).$$

This can also be written as

$$h^*\bar{Q}_0/\theta_0(Y\theta_0/\bar{Q}_0 - \theta_0),$$

where $\bar{Q}_0/\theta_0 = \exp(m_{\beta_0})$, and $h^*\bar{Q}_0/\theta_0$ is a function satisfying $E_0(h^*\bar{Q}_0/\theta_0 \mid W) = 0$. This latter representation can be used to prove that $S^*_{\beta_0,\theta,g_0}(O)$ has mean zero for all θ, thereby proving the double robustness of the efficient score as estimating function for β_0.

In the general case, for any $m_{\beta_0}(A,V)$, the clever covariate is given by

$$H^*_{\bar{Q}_0,g_0} = \frac{d}{d\beta}m_{\beta_0}(A,V) - \frac{E_0\left[\frac{\bar{Q}_0(A,W)}{1-\bar{Q}_0(A,W)}\frac{d}{d\beta}m_{\beta_0}(A,V)|W\right]}{E_0\left[\frac{\bar{Q}_0(A,W)}{1-\bar{Q}_0(A,W)}|W\right]}.$$

TMLE 2: Incorrectly assuming Y is a count of events and has a Poisson distribution

When we incorrectly assume that $P_0(Y|A,W)$ follows a Poisson distribution, the efficient score for the parameter of interest, β_0 in our semi-parametric multiplicative model is given by the following when we let $m_{\beta_0}(A,V) = \beta_0 A$:

$$S^*_{\bar{Q}_0,g_0}(O) = \left(A - \frac{E_0\left[Ae^{\beta_0 A}|W\right]}{E_0\left[e^{\beta_0 A}|W\right]} \right)(Y - \bar{Q}_0).$$

Then, as above, to find $r^*_{\bar{Q}_0,g_0}$ for this TMLE, we arrange that the score of the fluctuation model at $\epsilon = 0$ equals this efficient score. The clever covariate is then given by

$$H^*_{\bar{Q}_0,g_0} = A - \frac{E_0\left[Ae^{\beta_0 A}|W\right]}{E_0\left[e^{\beta_0 A}|W\right]}. \tag{4.4}$$

Detailed derivations for both the efficient score, $r^*_{\bar{Q}_0,g_0}$ and $H^*_{\bar{Q}_0,g_0}$ under the Poisson density assumption are also provided in the next chapter. In this case, the efficient score in the general case is given by

$$S^*_{\bar{Q}_0,g_0}(O) = \left(\frac{d}{d\beta_0} m_{\beta_0} - \frac{E_0\left[\frac{d}{d\beta_0} m_{\beta_0} e^{m_{\beta_0}}|W\right]}{E_0\left[e^{m_{\beta_0}}|W\right]} \right)(Y - \bar{Q}_0). \tag{4.5}$$

As above, we can again represent the efficient score as

$$S^*_{\bar{Q}_0,g_0}(O) = h^*(A \mid W)(Y - \bar{Q}_0),$$

but now

$$h^*(A \mid W) = \frac{d}{d\beta_0} m_{\beta_0} - \frac{E_0\left[\frac{d}{d\beta} m_{\beta_0} e^{m_{\beta_0}}|W\right]}{E_0\left[e^{m_{\beta_0}}|W\right]},$$

which again is a function satisfying $E_0(h^*(A \mid W)\exp(m_{\beta_0}) \mid W) = 0$, so that again we can prove that $S^*_{\beta_0,\theta,g_0}(O)$ has mean zero for all θ, proving the double robustness of the efficient score as estimating function for β_0 for this TMLE as well.

Here, the clever covariate in the general case for $m_{\beta_0}(A,V)$ is given by

$$H^*_{\bar{Q}_0,g_0} = \frac{d}{d\beta} m_{\beta_0}(A,V) - \left\{ \frac{E[\frac{d}{d\beta} m_{\beta_0}(A,V) e^{m_{\beta_0}(A,V)}|W]}{E[e^{m_{\beta_0}(A,V)}|W]} \right\}. \tag{4.6}$$

In addition to being DR, as mentioned earlier, this TMLE is an efficient estimator of β_0 in our assumed semi-parametric multiplicative model *if* Y is a count of events *and* the distribution for $P_0(Y|A,W)$ truly is Poisson. In this case, the parameter that the TMLE is estimating, when $m_{\beta_0} = \beta_0 A$ for example, can be interpreted as the change in the log

conditional *incidence rate* associated with a unit change in A, rather than the change in the log conditional relative risk. However, we can apply this TMLE to estimate the change in log conditional relative risk as well - when our outcome is truly binary and therefore $P_0(Y|A,W)$ cannot have a Poisson distribution. *For this target parameter*, this second TMLE is no longer efficient. However, we can still achieve asymptotic linearity and proper inference. Therefore, we opt to implement this TMLE as an estimator of β_0 with binary Y. The TMLE is computationally stable, unlike the first TMLE, which very often suffers from non-convergence due to it's reliance on log-binomial regression. Going forward, we provide a step-by-step implementation for this practical TMLE and we demonstrate its performance in simulations. But first, we briefly discuss how we can adapt the TMLE's when the data come from an independent case-control, rather than from a prospective, sampling design.

4.4.3 Adapting the TMLE's for Independent Case-Control Sampling

Case-control studies are commonly used to analyze parameters for binary outcomes when the probability of an event (referred to as the prevalence probability) is very small in the population of interest. Case-control data are biased in that the number of cases in the data is disproportionate to the number in the sampled population. This allows for a sufficient number of cases to be obtained without extensive sampling. Case-control sampling can also result increased efficiency relative to prospective sampling. Both TMLE's discussed above can easily to be adapted to be applied to data from an independent case-control design.

With such a design, the experimental unit is a *cluster* of observations consisting of one case and J controls. As described in van der Laan [2008] and Rose and van der Laan [2008], (W_1, A_1) are first sampled from the conditional distribution of (W, A) given $Y = 1$. Next, (W_0^j, A_0^j) are sampled from (W, A) given $Y = 0$ for $j = 1, \ldots, J$. Therefore, the observed data structure for independent case-control sampling is defined as

$$O = ((W_1, A_1), (W_0^j, A_0^j), Y) \sim P_0,$$

such that $(W_1, A_1) \equiv (W, A|Y = 1)$ and $(W_0^j, A_0^j) \equiv (W, A|Y = 0)$. The underlying experimental unit is represented by $O^* = (W, A, Y) \sim P_0^*$.

The adaptations to the above TMLE's when we have this case-control data structure are straight-forward. First, we estimate β_n^0 using the case-control weighting described in van der Laan [2008] and Rose and van der Laan [2008]. In van der Laan [2008], it is shown that with such a weighting scheme, any estimation method developed for prospective sampling can be mapped into an estimation method based on case-control sampling. The weighting method requires knowledge of the incidence probability in the sampled population, q_0. If q_0 is unknown, it may be estimated as reliably as possible from existing studies.

As above, for the TMLE's applied to prospective sampling, we first estimate θ_0, using a single data-adaptive estimator or super learner, forcing A into the model, but now cases are

weighted by q_0 and controls are weighted by $(1-q_0)/J$. Then as before, we set $A=0$ for all observations to get θ_n. We proceed with a parametric regression model (either log-binomial or Poisson), again with case-control weighting, to obtain, β_n^0 and $\bar{Q}_n^0$. Next, the targeted maximum likelihood fluctuation step follows the same general approach. The only difference is that observations' assigned weights are used with each update. Also, in order to obtain the correct inference (i.e. correct standard error estimates), we need to use the correct version of the efficient influence curve so that the relationships in the data structure are taken into account. This is presented in the next section, a step-by-step guide to implementing the Poisson-derived, "practical" TMLE in either a prospective or case-control sample.

4.5 Step-by-Step Implementation

In this section, we provide a step-by-step description of how to implement the Poisson-derived, "practical" TMLE to estimate β_0 in the assumed semi-parametric multiplicative regression model for a binary Y, as shown in (4.1).

(1) If data are from an independent case-control design, assign case and control weights.

Let $q_0 \equiv P_0(Y=1)$, and assign weights q_0 to the cases and weights $(1-q_0)/J$ to the J controls. If q_0 is unknown, it may be estimated as reliably as possible from existing studies.

(2) Define $m_{\beta_0}(A,V)$.

Based on substantive knowledge, the analyst needs to decide on the best model to assume for $m_{\beta_0}(A,V)$ (e.g. $\beta_0(A)$ or $\beta_0'[A,A:V]$).

(3) Estimate $\bar{Q}_0(A,W)$.

To obtain the initial estimator of $\bar{Q}_0$, we first first need an estimator of θ_0. To do so, we first obtain a general form estimator of $\bar{Q}_0$, using a data adaptive algorithm, preferably super learner, forcing A into the model. We then set $A=0$ for all observations in order to get θ_n^0. Then, we fit $\log\bar{Q}_0(A,W) = m_{\beta_0}(A,V) + \log\theta_0(W)$ using a Poisson regression model to obtain the initial estimator $\bar{Q}_n^0$. If we have independent case-control data, we weigh the observations by their assigned weights q_0 or $(1-q_0)/J$.

(4) If unknown, estimate $g_0(A|W)$.

In clinical trials and other experiments, A is randomized and g_0 is known. If g_0 is unknown, then when A is binary or categorical, estimation is straight-forward. Ideally, we would use a data-adaptive procedure or super learning. If A is continuous, we can

skip directly estimating g_0 and instead use a data-adaptive procedure to obtain both the numerator and denominator of the second term of the clever covariate given in (4.4) and (4.6), both which incorporate g_n.

(5) Calculate the clever covariate estimator, $H^*_{\bar{Q}^0_n,g_n}(W)$.

Referring to (4.6), $H^*_{\bar{Q}^0_n,g_n}$ is based on observed values of A, W, $\bar{Q}^0_n$ and g_n (or g_0 if known). For example, if we have $m_{\beta_0}(A,V) = \beta_0 A$ and $A \in [0,1]$, the initial estimator of the clever covariate is given by

$$H^*_{\bar{Q}^0_n,g_n} = A - \frac{e^{\beta^0_n} g_n(1,W)}{e^{\beta^0_n} g_n(1|W) + g_n(0|W)}.$$

Or, if A is a non-ordered categorical variable with J values, $H^*_{\bar{Q}^0_n,g_n}$ is a j dimensional vector such that the j^{th} entry is given by

$$H^*_{\bar{Q}^0_n,g_n,j} = A - \frac{e^{\beta^0_n I(A=j)} g_n(A=j|W)}{e^{\beta^0_n I(A=j)} g_n(A=j|W) + (1 - g_n(A=j|W))}.$$

As noted in Step (4), when A is continuous, the numerator and denominator of the second term in $H^*_{\bar{Q}^0_n,g_n}$, can each be estimated with machine learning. We do not necessarily recommend super learning here, as the TMLE may require many iterations and super learning can be too computationally intensive.

(6) Update $\bar{Q}^0_n(A,W)$.

We update $\bar{Q}^0_n$ by fitting the following regression model:

$$\log\bar{Q}^1_n(A,W) = \beta^0_n A + \log\theta^0_n + \epsilon^1 H^*_{\bar{Q}^0_n,g_n}.$$

Therefore, to implement, we can use standard regression software for Poisson models, regressing Y on $H^*_{\bar{Q}^0_n,g_n}$ with no intercept and and with $\log\bar{Q}^0_n$ as an offset. The updated parameter estimate is given by $\beta^1_n = \beta^0_n + \epsilon^1_n$. If we have independent case-control data, we use weighted maximum likelihood using the assigned weights from Step (1).

We iterate this process (Steps 5 and 6) until $\epsilon^k_n \approx 0$. After iterating k times, the estimator of our parameter is given by $\beta^*_n = \beta^k_n$. Therefore, the TMLE β^*_n is an update of the original estimate β^0_n, correcting for bias due to residual confounding and non-targeting.

(7) Calculate standard errors.

If we assume that g_n is consistent, we can estimate the variance of β^*_n with the empirical variance of the efficient influence curve. The efficient influence curve is given by

$$D^*_{\bar{Q}_0,g_0}(O) = -E\left[\frac{d}{d\beta_0}S^*_{\bar{Q}_0,g_0}(O)\right]^{-1} S^*_{\bar{Q}_0,g_0}(O),$$

where $S^*_{\bar{Q}_0,g_0}(O)$ is the efficient score given in (4.5).

For a case-control design, the efficient influence curve for the cluster observation O is given by the following if, for example, $j = 1$:

$$\begin{aligned} D^*_{\bar{Q}_0,g_0}(O) &= [q_0\frac{d}{d\beta_0}E_0S^*_{\bar{Q}_0,g_0}(1,A_1,W_1) + (1-q_0)\frac{d}{d\beta_0}E_0S^*_{\bar{Q}_0,g_0}(0,A_0,W_0)]^{-1} \\ &\quad [q_0S^*_{\bar{Q}_0,g_0}(1,A_1,W_1) + (1-q_0)S^*_{\bar{Q}_0,g_0}(0,A_0,W_0)]. \end{aligned}$$

Using the empirical variance of the appropriate efficient influence curve, we can then calculate standard errors, p-values and confidence intervals.

4.6 Simulations

In this section, we assess the properties of the Poisson-derived TMLE of β_0 with simulations that cover a range of scenarios seen in actual data sets. With these simulations, we demonstrate the double robustness properties of the Poisson-derived, practical TMLE and differences in variability when we estimate $\bar{Q}_0(A,W)$ and/or $g_0(A|W)$ consistently or with super learner. We compare our results to those from common estimators in the literature - for a prospective design, those obtained from parametric log-binomial and Poisson regression models and for an independent case-control design, those obtained from a parametric logistic regression model (approximating conditional relative risk with the conditional odds ratio). Overviews and comparisons of the methods for a prospective design can be found in papers by L. McNutt and Hafner [2003], Barros and Hirakata [2003] and Lumley and Ma [2006], and a discussion of using logistic regression to approximate conditional relative risk can be found in Hogue et al. [1983],Greenland [2003]. We also compare the TMLE results to those obtained from an original fit of the semi-parametric model, using super learner to estimate θ_0. To evaluate the performance of all estimators, we focus on bias, variance, mean squared error (MSE) and confidence intervals.

We show that the relative performance of the TMLE, when compared to the other estimators, depends partly on the degree of practical positivity violations, that is within strata defined by W, the extent to which values of g_n are bounded away from 0 and 1. When A is binary or categorical and is perfectly randomized, this is not an issue. The common methods in the literature perform well under this scenario. However, in observational studies, analysts often have the challenge that some values of $g_n(A|W)$ are very small, particularly when there are many covariates in W, some covariates are continuous, and/or when A is continuous.

Different estimators are affected by positivity violations in different ways. In the following simulations, we demonstrate the relative performance of the practical TMLE of β_0, for both a

binary and a continuous A, when (1) A is perfectly randomized, (2) the relationship between A and Y is confounded by W but there are no positivity violations and (3) there are extreme positivity violations.

In all simulations, $Y \in \{0, 1\}$ is a binary outcome such as an indicator of disease status. Also in all of the simulations, W is a vector of five covariates, which were generated as follows:

$$\begin{aligned} W_1 &\sim Binom(1, 0.3) \\ W_2 &\sim Binom(1, 0.65) \\ W_3 &\sim N(0, 2) \\ W_4 &\sim N(100, 10) \\ W_5 &\sim N(1, 0.3). \end{aligned}$$

$\bar{Q}_0(A, W)$ is given by

$$\bar{Q}_0(A, W) = e^{-0.1A} e^{I + 0.1W_3 + 0.02W_2W_3 - 0.01W_1W_4 - 0.02W_5}, \tag{4.7}$$

where I takes the following values for the various simulations:

	Binary A			Continuous A		
Design	Simulation 1	Simulation 2	Simulation 3	Simulation 1	Simulation 2	Simulation 3
Prospective	-0.8	-0.8	-1.0	-0.8	-0.4	-1.4
Case-control	-4.0	-4.0	-4.0	-4.0	-4.0	-4.0

As (4.7) shows, $\beta_0 = -0.1$ in all simulations. Also, (4.7) shows that $m_{\beta_0} = \beta_0 A$, so we have assumed there are no effect modifiers. For the case-control data simulations, we followed case-control sampling as described in Section 4.4.3 with $j = 1$, so that for each case, there is one control.

4.6.1 Simulations for Binary A

For a binary A, we consider the following three conditional distributions for $g_0(A|W)$:

1. For the first simulation, A is perfectly randomized such that $g_0(A|W) = 0.5$. When $g_0(A|W)$ is misspecified for this simulation, $g_n(A|W) = 0.6$.

2. For the second simulation, A is dependent on W such that

$$g_0(A|W) = \frac{1}{1 + exp(-(0.1W_3))}.$$

With this mechanism for exposure, the correlation between A and W_3 is 0.10, and values of $g_0(A|W)$ range from 0.28 to 0.73, with a median of 0.49. Therefore, we do not have positivity violations. For this simulation, when $g_n(A|W)$ is misspecified, the estimator depends only on W_1.

3. For the third simulation, A is again dependent on W; but now we have

$$g_0(A|W) = \frac{1}{1 + exp(-(1.0W_3))}.$$

This mechanism for exposure leads to positivity violations because $g_0(A|W) \in [5.4 \times 10^{-5}, 1.0]$. The median value is 0.54. The correlation between A and W_3 is now 0.61. Misspecification of $g_n(A|W)$ again occurs by having the estimator only depend on W_1.

4.6.2 Simulations for Continuous A

For continuous A, we again varied $g_0(A|W)$ three ways:

1. For the first simulation, A is not dependent on W. It is normally distributed such that $A \sim N(1, 0.6)$.

2. For the second simulation, A is dependent on W such that $A \sim N(1, 0.6) + 0.1W_3$. In this simulation, the correlation between A and W_3 is -0.1.

3. For the third simulation, A is dependent on W such that $A \sim N(0, 0.6) - 0.8W_3$. The correlation between A and W_3 in this simulation is -0.6.

For all simulations, we generated 1000 samples of size 1000. All data were generated and all estimators were implemented using R (Team [2010]).

4.6.3 Prospective Sample Simulation Results

Tables 4.1 and 4.2 present results for estimating β_0 from a prospective sampling design, when A is binary and when A is continuous. For a continuous A, we estimated the numerator and denominator of the clever covariate using the lars package in R (Efron et al. [2003]). The first column of the tables presents the initial substitution estimator, β_n^0, based on the initial estimate of $\bar{Q}_0$. The second column presents the TMLE, β_n^*, obtained by substitution after k iterations of updating $\bar{Q}_n^0$ to obtain $\bar{Q}_n^*$. The subsequent columns provide the estimated bias, mean squared error (MSE) and empirical variance, calculated from 1000 samples. We also include the mean of the variance estimates calculated from the empirical variance of efficient influence curve divided by the sample size of 1000. Finally, the last column shows the coverage probability (CP), or the percentage of the time that the estimated 95% confidence interval contains the true value of $\beta_0 = -0.1$.

Each panel in the tables corresponds to the simulations described above, and the rows in each panel indicate the specification of the estimators of $\bar{Q}_0$ and g_0, on which the TMLE is based. For example, "Qcgc" indicates that the correct terms were included when estimating both $\bar{Q}_0$ and g_0. "Qcgm" indicates that the estimator of g_0 was misspecified as described above, while the estimator of $\bar{Q}_0$ included the correct terms; and "Qgmc" indicates that the estimator of $\bar{Q}_0$ was misspecified as described above, while the correct terms were included when estimating g_0. Finally "Qslgsl" indicates that the super learner was used for the estimators of both $\bar{Q}_0$ and g_0.

Tables 4.1 and 4.2 illustrate the properties we expect to see for the TMLE of β_0:

- The TMLE is double-robust. The finite-sample bias is close to zero if the estimator of either $\bar{Q}_0$ or g_0 is consistent. We achieve this result even under substantial confounding and extreme violations of positivity in Simulation 3.

- When the estimator of g_0 is consistent, the variance estimate obtained from the empirical variance of the efficient influence curve is approximately equal to the variance of the 1000 TMLEs and the coverage probability is approximately 95%. When the estimator of g_n is inconsistent, this variance estimate is asymptotically conservative.

- Using the super learner to estimate both $\bar{Q}_0$ and g_0 provides robust estimates of either $\bar{Q}_0$ or g_0 so that we achieve comparable bias and variance as obtained when correctly specifying the models for $\bar{Q}_0$ and/or g_0.

Tables 4.3 and 4.4 compare the performance of the TMLE of β_0 to the common estimators in the literature when the initial working model for $\bar{Q}_0(A, W)$ is *incorrect*. All of the estimators in the literature will perform well when the parametric models on which they rely are correctly specified. However, we are very doubtful that anyone can ever specify a parametric model correctly. Therefore, we present comparisons under a more realistic scenario.

Within each panel for each simulation, the first two rows present results (bias, variance and MSE) for the common methods in the literature - using log binomial and Poisson regres-

Table 4.1: Performance of Poisson-derived TMLE, binary A, by simulation for prospective sample

	β_n^0	β_n^*	Bias	MSE	Var(β_n^*)	Var(IC_{eff})/n	CP
Simulation 1							
Qcgc	-0.102	-0.102	-0.002	0.006	0.006	0.007	0.964
Qcgw	-0.102	-0.102	-0.002	0.006	0.006	0.011	0.990
Qwgc	-0.101	-0.101	-0.001	0.006	0.006	0.007	0.966
Qslgsl	-0.090	-0.102	-0.002	0.006	0.006	0.007	0.960
Simulation 2							
Qcgc	-0.103	-0.103	-0.003	0.007	0.007	0.007	0.950
Qcgw	-0.103	-0.103	-0.003	0.007	0.007	0.007	0.952
Qwgc	-0.057	-0.102	-0.002	0.007	0.007	0.007	0.944
Qslgsl	-0.088	-0.101	-0.001	0.007	0.007	0.007	0.948
Simulation 3							
Qcgc	-0.111	-0.111	-0.011	0.017	0.017	0.016	0.950
Qcgw	-0.111	-0.111	-0.011	0.016	0.016	0.008	0.816
Qwgc	0.169	-0.109	-0.009	0.017	0.017	0.016	0.944
Qslgsl	-0.091	-0.109	-0.009	0.017	0.017	0.016	0.940

Table 4.2: Performance of Poisson-derived TMLE, continuous A, by simulation for prospective sample

	β_n^0	β_n^*	Bias	MSE	Var(β_n^*)	Var(IC_{eff})/n	CP
Simulation 1							
Qcgc	-0.099	-0.099	0.001	0.005	0.005	0.005	0.946
Qcgw	-0.099	-0.099	0.001	0.005	0.005	0.005	0.946
Qwgc	-0.098	-0.098	0.002	0.005	0.005	0.005	0.950
Qslgsl	-0.089	-0.099	0.001	0.005	0.005	0.005	0.950
Simulation 2							
Qcgc	-0.098	-0.098	0.002	0.005	0.005	0.005	0.956
Qcgw	-0.098	-0.098	0.002	0.005	0.005	0.005	0.956
Qwgc	0.016	-0.099	0.001	0.005	0.005	0.005	0.956
Qslgsl	-0.089	-0.099	0.001	0.005	0.005	0.005	0.958
Simulation 3							
Qcgc	-0.102	-0.103	-0.003	0.003	0.003	0.065	0.956
Qcgw	-0.102	-0.103	-0.003	0.003	0.003	0.065	0.956
Qwgc	0.025	-0.103	-0.003	0.003	0.003	0.004	0.958
Qslgsl	-0.096	-0.105	-0.005	0.003	0.003	0.003	0.948

Table 4.3: Relative performance of Poisson-derived TMLE, binary A, prospective sample

	Bias	Var	MSE
Simulation 1			
Log Binomial, incorrect	-0.004	0.007	0.007
Poisson, incorrect	-0.004	0.007	0.007
β_n^0, incorrect	-0.001	0.006	0.006
β_n^0, SL	0.010	0.008	0.008
β_n^* Qwgc	-0.001	0.006	0.006
β_n^* Qslgsl	-0.002	0.006	0.006
Simulation 2			
Log Binomial, incorrect	0.046	0.007	0.009
Poisson, incorrect	0.047	0.007	0.009
β_n^0, incorrect	0.043	0.007	0.009
β_n^0, SL	0.012	0.009	0.009
β_n^* Qwgc	-0.002	0.007	0.007
β_n^* Qslgsl	-0.001	0.007	0.007
Simulation 3			
Log Binomial, incorrect	0.277	0.010	0.087
Poisson, incorrect	0.277	0.010	0.087
β_n^0, incorrect	0.269	0.010	0.083
β_n^0, SL	0.009	0.016	0.016
β_n^* Qwgc	-0.009	0.017	0.017
β_n^* Qslgsl	-0.009	0.017	0.017

sion to estimate W-adjusted relative risk. The third and fourth rows present results for the initial estimate of β_0, β_n^0, when $\bar{Q}_n^0(A, W)$ is incorrectly specified and when it is estimated by super learning. The last two rows then present results for the TMLE, β_n^*, when $\bar{Q}_n^0(A, W)$ is incorrectly specified and when it is estimated by super learning.

Figures 4.1 and 4.2 also compare the performance of TMLE to other relative risk estimators The following summarizes key observations from both the tables and figure:

- In a randomized trial, as demonstrated in Simulation 1, all estimators perform comparably well, as expected, for both binary and continuous A.
- As the relationship between the true confounder and A increases in Simulations 2 and 3, the estimators utilizing on log-binomial regression and Poisson regression are increasingly biased, while the variance (of the 1000 sample estimates of β_0) remains at the same or similar level (for binary A) or decreases (for continuous A).
- The TMLE's of β_0 achieve the lowest MSE in both simulations with confounding (Simulations 2 and 3). We see a small trade-off in variance for removal of bias.
- Even with positivity violations, the TMLE's are relatively robust.

Table 4.4: Relative performance of Poisson-derived TMLE, continuous A, prospective sample

	Bias	Var	MSE
Simulation 1			
Log Binomial, incorrect	0.000	0.004	0.004
Poisson, incorrect	-0.002	0.005	0.005
β_n^0, incorrect	0.002	0.005	0.005
β_n^0, SL	0.011	0.007	0.007
β_n^* Qwgc	0.002	0.005	0.005
β_n^* Qslgsl	0.001	0.005	0.005
Simulation 2			
Log Binomial, incorrect	0.111	0.004	0.017
Poisson, incorrect	0.110	0.004	0.016
β_n^0, incorrect	0.116	0.005	0.018
β_n^0, SL	0.011	0.007	0.007
β_n^* Qwgc	0.001	0.005	0.005
β_n^* Qslgsl	0.001	0.005	0.005
Simulation 3			
Log Binomial, incorrect	0.120	0.000	0.015
Poisson, incorrect	0.123	0.000	0.015
β_n^0, incorrect	0.125	0.000	0.016
β_n^0, SL	0.004	0.003	0.003
β_n^* Qwgc	-0.003	0.003	0.003
β_n^* Qslgsl	-0.005	0.003	0.003

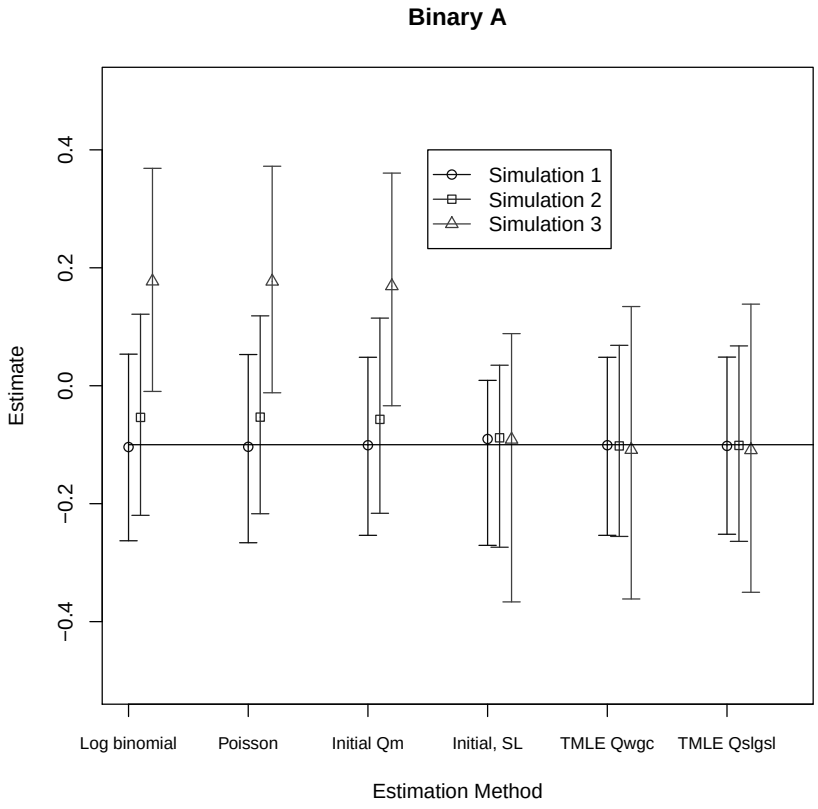


Figure 4.1: Estimates and 95% confidence intervals by method, binary A, prospective sample

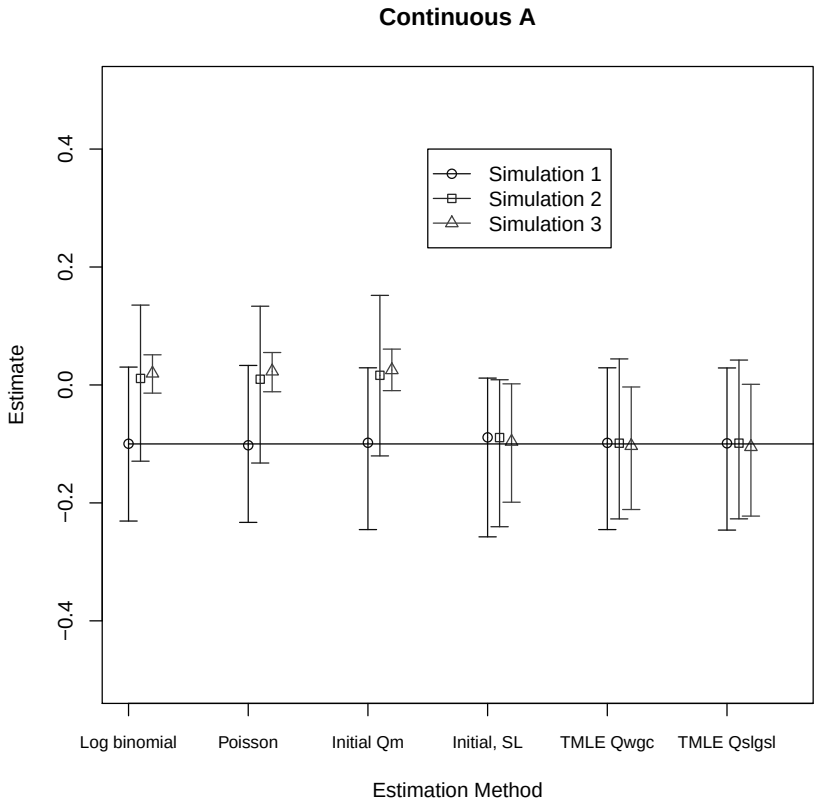


Figure 4.2: Estimates and 95% confidence intervals by method, continuous A, prospective sample

Table 4.5: Performance of Poisson-derived TMLE, binary A, by simulation for case-control sample

	β_n^0	β_n^*	Bias	MSE	Var(β_n^*)	Var(IC_{eff})/n	CP
Simulation 1							
Qcgc	-0.096	-0.096	0.004	0.018	0.017	0.017	0.945
Qcgw	-0.096	-0.096	0.004	0.018	0.017	0.027	0.984
Qwgc	-0.097	-0.097	0.003	0.017	0.017	0.017	0.950
Qslgsl	-0.017	-0.096	0.004	0.016	0.016	0.015	0.944
Simulation 2							
Qcgc	-0.111	-0.111	-0.011	0.020	0.020	0.017	0.932
Qcgw	-0.111	-0.111	-0.011	0.020	0.020	0.018	0.935
Qwgc	-0.064	-0.109	-0.009	0.019	0.019	0.017	0.939
Qslgsl	-0.020	-0.099	0.001	0.017	0.017	0.015	0.939
Simulation 3							
Qcgc	-0.109	-0.108	-0.008	0.031	0.031	0.028	0.931
Qcgw	-0.109	-0.109	-0.009	0.029	0.029	0.014	0.833
Qwgc	0.172	-0.104	-0.004	0.030	0.030	0.028	0.933
Qslgsl	-0.018	-0.101	-0.001	0.029	0.029	0.025	0.931

4.6.4 Case-Control Sample Simulation Results

Tables 4.5 and 4.6 present the same results when estimating β_0 from a case-control sample. The results have the same properties that we observed for prospective sampling.

Tables 4.7 and 4.8 and Figures 4.3 and 4.3 show the performance of the practical, Poisson-derived TMLE of β_0 compared to using logistic regression and then converting odd-ratio parameters to relative risk parameters, as the odds ratio approximates the relative risk when the prevalence probability is close zero (Hogue et al. [1983],Greenland [2003]). Bias in this approximation increases as the true prevalence probability increases. As above we compare results when the working parametric model for $\bar{Q}_0(A, W)$ is *misspecified*.

As we saw in the results from prospective sampling, all estimators perform well when A is randomized in Simulation 1. In Simulations 2 and 3, with more confounding and with theoretical ETA violations, the TMLE's perform much better than using logistic regression to estimate an odds ratio and then using that odds ratio as an estimate of relative risk. For both both binary and continuous A, the TMLE's achieve far less bias and better MSE.

Table 4.6: Performance of Poisson-derived TMLE, continuous A, by simulation for case-control sample

	β_n^0	β_n^*	Bias	MSE	Var(β_n^*)	Var(IC_{eff})/n	CP
Simulation 1							
Qcgc	-0.100	-0.100	0.000	0.013	0.013	0.011	0.912
Qcgw	-0.100	-0.100	0.000	0.013	0.013	0.011	0.912
Qwgc	-0.099	-0.099	0.001	0.013	0.013	0.011	0.919
Qslgsl	-0.019	-0.101	-0.001	0.012	0.012	0.010	0.919
Simulation 2							
Qcgc	-0.100	-0.100	0.000	0.012	0.012	0.011	0.931
Qcgw	-0.100	-0.100	0.000	0.012	0.012	0.011	0.931
Qwgc	0.014	-0.098	0.002	0.011	0.011	0.011	0.937
Qslgsl	-0.018	-0.098	0.002	0.011	0.011	0.010	0.942
Simulation 3							
Qcgc	-0.097	-0.102	-0.002	0.016	0.016	0.020	0.949
Qcgw	-0.097	-0.102	-0.002	0.016	0.016	0.020	0.949
Qwgc	0.023	-0.102	-0.002	0.016	0.016	0.323	0.946
Qslgsl	-0.017	-0.102	-0.002	0.016	0.016	1.858	0.946

Table 4.7: Relative performance of Poisson-derived TMLE, binary A, case-control sample

	Bias	Var	MSE
Simulation 1			
Logistic, incorrect	0.002	0.017	0.017
β_n^0, incorrect	0.003	0.017	0.017
β_n^0, SL	0.083	0.001	0.008
β_n^* Qwgc	0.003	0.017	0.017
β_n^* Qslgsl	0.004	0.016	0.016
Simulation 2			
Logistic, incorrect	0.037	0.019	0.020
β_n^0, incorrect	0.036	0.019	0.020
β_n^0, SL	0.080	0.001	0.007
β_n^* Qwgc	-0.009	0.019	0.019
β_n^* Qslgsl	0.001	0.017	0.017
Simulation 3			
Logistic, incorrect	0.275	0.017	0.093
β_n^0, incorrect	0.272	0.017	0.091
β_n^0, SL	0.082	0.001	0.008
β_n^* Qwgc	-0.004	0.030	0.030
β_n^* Qslgsl	-0.001	0.029	0.029

Table 4.8: Relative performance of Poisson-derived TMLE, continuous A, case-control sample

	Bias	Var	MSE
Simulation 1			
Logistic, incorrect	0.000	0.013	0.013
β_n^0, incorrect	0.001	0.013	0.013
β_n^0, SL	0.081	0.001	0.007
β_n^* Qwgc	0.001	0.013	0.013
β_n^* Qslgsl	-0.001	0.012	0.012
Simulation 2			
Logistic, incorrect	0.114	0.010	0.023
β_n^0, incorrect	0.114	0.010	0.024
β_n^0, SL	0.082	0.001	0.007
β_n^* Qwgc	0.002	0.011	0.011
β_n^* Qslgsl	0.002	0.011	0.011
Simulation 3			
Logistic, incorrect	0.124	0.001	0.017
β_n^0, incorrect	0.123	0.002	0.017
β_n^0, SL	0.083	0.001	0.007
β_n^* Qwgc	-0.002	0.016	0.016
β_n^* Qslgsl	-0.002	0.016	0.016

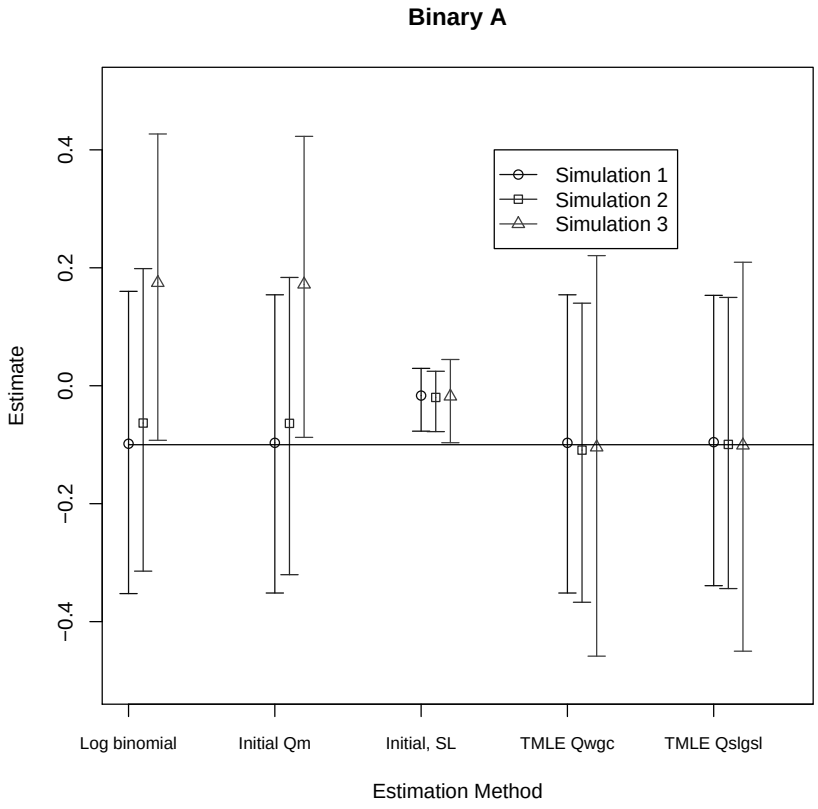


Figure 4.3: Estimates and 95% confidence intervals by method, binary A, case-control sample

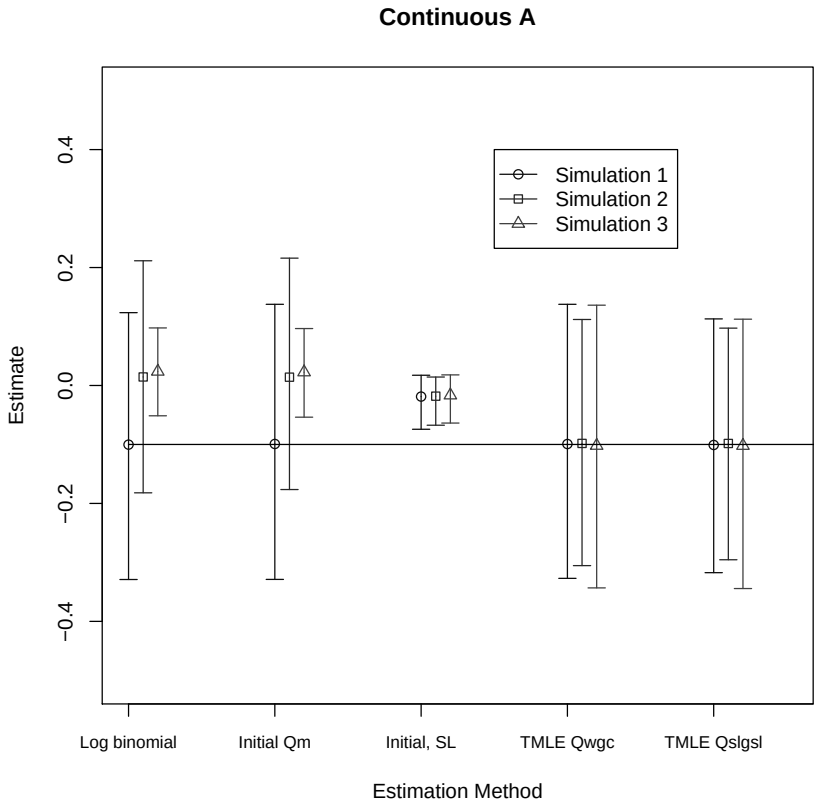


Figure 4.4: Estimates and 95% confidence intervals by method, continuous A, case-control sample

4.7 Discussion

This chapter introduced two TMLE's for a parameter that relates an exposure to conditional relative risk, a common objective in medical and epidemiology studies when the outcome is binary. The TMLE's were developed under a semi-parametric multiplicative model, which provides more flexibility than a fully parametric model, while still providing an interpretable parameter of interest. In practice, semi-parametric modeld can of course be misspecified. However, semi-parametric models are particularly attractive when there is interest in effect of a continuous exposure since nonparametric model-based estimators of the effects of continuous exposures tend to be unstable. Note that an alternative approach is to define parameters in nonparametric models that are extensions of the parameters as defined by semi-parametric regression models, so that the target parameter remains defined outside the semi-parametric model. Such work is in progress.

The first of the two TMLE's in this chapter is the "correct" TMLE for a binary outcome and was derived by correctly assuming a binomial density. The second TMLE is the correct TMLE for a count outcome with a Poisson distribution and was derived under this assumption. The Poisson assumption is always wrong when the outcome is truly binary, but we can apply this second TMLE to binary data. We refer to this as the "practical" TMLE because it is does not suffer from convergence problems like the first, correct TMLE. When applied to data with a binary outcome, this practical TMLE is not efficient, but it does remain DR, asymptotically linear and achieves the correct inference. If the outcome were instead a count of events and followed a Poisson distribution, this second TMLE would be an efficient estimator of β_0 in the model (and the interpretation of β_0 would change).

Unlike other some other TMLE's, including those in Chapters 1 and 2, the TMLE's presented in this chapter do not depend on inverse probability weighting, and therefore should be relatively robust to positivity violations. We demonstrated this property for the practical, Poisson-derived TMLE with simulation studies. We also compared this TMLE to common parametric methods in the literature and illustrated the strong performance under various degrees of positivity violations. We also saw the superior performance of the TMLE as confounding increased. We confirmed these findings for both prospective and case-control samples.

In sum, this chapter has illustrated the value of introducing model assumptions in order to address positivity violations. Because we relied on a semi-parametric model rather than a parametric model, we have maximized flexibility. We have shown that with a TMLE, we can obtain robust estimates of a parameter that relates exposure to changes in conditional relative risk, even under strong confounding and positivity violations. We demonstrated the practical, Poisson-derived TMLE's DR properties in both prospective and case-control samples, illustrated that it achieves proper inference and showed its improved performance over existing methods. The next chapter provides more methodological details for both TMLE's.

Chapter 5

Constructing the Efficient Score and Clever Covariate for Targeted Maximum Likelihood Estimators of Conditional Relative Risk Parameters in a Semi-parametric Regression Model

5.1 Introduction

This chapter delves into the methodological details of the TMLE's introduced in Chapter 3, as well as in Tuglus et al. [2011]. The TMLE's estimate parameters that relate log conditional relative risk to exposure as defined by a semi-parametric multiplicative regression model. One TMLE correctly assumes that the binary outcome (i.e. disease or no disease) has a binomial distribution, and one TMLE treats the binary outcome as a count of events and therefore incorrectly assumes a Poisson distribution. The first of these TMLE's is an efficient estimator of the parameter of interest but is unstable due to convergence problems with the log-binomial model, which is used for estimation. The second of these TMLE's is an efficient estimator of change in the *log conditional incidence rate* associated with a unit change in exposure, the parameter of interest when the outcome is truly a count of events and follows a Poisson distribution. However, this second TMLE can also and has been applied to a binary outcome, estimating our first parameter of interest. In this case, the TMLE is no longer efficient, but it does achieve stability. It also remains DR - that is, the efficient score equation solved by this second TMLE is unbiased DR estimating equation for the parameter of interest in the conditional mean model, which does not assume a Poisson distribution.

Consequently, we obtain consistency and correct inference.

The previous chapter and Tuglus et al. [2011] present some basic theoretical features needed to construct both the "correct" TMLE (for a binary outcome) and the "practical" TMLE (for a count outcome applied to a binary outcome) - the efficient scores and efficient influence curves for the two different targeted parameters, as well as the parametric fluctuation submodels for the TMLE step, including the so called clever covariates that define the submodels. This chapter provides important methodological details for deriving these features. Specifically, this chapter first provides the theoretical derivation of the efficient score and efficient influence curve for our primary parameter of interest, change in log relative risk associated with a unit change in exposure. We also present the correct parametric submodel through the initial estimator of the density that has score equal to the efficient score. We thus derive the clever covariate needed for the fluctuation step. This results in an efficient, but unstable TMLE. This chapter also considers the case in which the outcome truly is a count of events. For the corresponding parameter of interest, the change in log conditional incidence rate associated with a unit change in exposure, we present the efficient score and efficient influence curve. Rather than carry out the entire derivation, we include a proof that the efficient score and efficient influence curve presented are indeed those. We then derive the clever covariate for this TMLE. We also discuss the properties of this TMLE when applied to our original parameter of interest, when the outcome is truly binary.

After a brief review in Section 5.2 of the data structure (focusing only on a prospective point-treatment sample), the semi-parametric regression model, the parameter implied by the model and the parametric fluctuation model, all as laid out in Chapter 3, the remainder of this chapter is organized as follows. In Section 5.3, we focus on the the "correct" TMLE for the relative risk parameter, and in Section 5.4, we focus on the TMLE for the incidence rate parameter, while commenting on applying it as a "practical" TMLE for the relative risk parameter, as we did in Chapter 3. The current chapter concludes with a summary in Section 5.5.

5.2 Overview of Data Structure, Semi-parametric Regression Model, Parameter and Parametric Fluctuation Model

As in Chapter 3, we consider n independent and identically distributed (i.i.d.) observations of $O = (W, A, Y) \sim P_0 \in \mathcal{M}$, where W is a vector of baseline covariates, A is an exposure of interest, and $Y = \{0, 1\}$ is a binary outcome (although also considered a count outcome for our second TMLE in Section 5.4). P_0 denotes the true distribution of O, from which all subjects are sampled, and P_0 is an element of the statistical model $\mathcal{M}$. We define the statistical model by our assumed semi-parametric multiplicative regression model

$$\bar{Q}_0(A,W) = e^{m_{\beta_0}(A,V)}\theta_0(W), \tag{5.1}$$

where $\bar{Q}_0(A,W) \equiv P_0(Y=1 \mid A,W)$ (or in the case of a count outcome, $\bar{Q}_0(A,W) \equiv E_0(Y \mid A,W)$), $m_{\beta_0}(A,V)$ is a specified function of A and effect modifiers $V \subset W$, and $\theta_0(W) \equiv P_0(Y=1 \mid A=0,W)$ (or in the case of count data, $\theta_0(W) \equiv E_0(Y \mid A=0,W)$). We focus on the case in which $m_{\beta_0}(A,V) = \beta_0 A$. Therefore, our parameter of interest when Y is binary, β_0, is equivalent to the change in the log conditional relative risk associated with a unit increase of exposure A on outcome Y. When Y is considered a count of events, the true parameter of interest, β_0, becomes the log conditional incidence rate associated with a unit increase in exposure A.

We obtain β_n^0, our initial estimate of β_0, by first estimating θ_0. We do this by using a pre-specified data-adaptive estimator or, preferably, an ensemble data-adaptive estimators such as super learner (van der Laan et al. [2007]), forcing A into the model to be selected. We then set $A=0$ for all observations to get predicted values for an estimator θ_n^0 of θ_0. Then to obtain β_n^0, we can use the parametric regression model corresponding to our assumption about the distribution of Y (i.e. log-binomial or Poisson regression). With θ_n^0 and β_n^0, we have the initial estimator $\bar{Q}_n^0$ of $\bar{Q}_0$.

For the targeted maximum likelihood step, we fluctuate $\bar{Q}_n^0$ (which we write as $\bar{Q}_{\beta_n^0,\theta_n^0}$ to emphasize the reliance on β_n^0 and θ_n^0) to be tailored to estimation of the parameter of interest, β_0. The general parametric fluctuation submodel (without making any particular parametric assumptions about the distribution of Y) is given by

$$\log\bar{Q}_{\beta_n^0,\theta_n^0}(\epsilon)(A,W) = (\beta_n^0+\epsilon)A + \log\theta_n^0(\epsilon)(W),$$

where $\theta_n^0(\epsilon)(W) = \theta_n^0(W)\exp(\epsilon r^*_{\bar{Q}_n^0,g_n}(W))$. The function $r^*_{\bar{Q}_n^0,g_n}(W)$ depends on $\bar{Q}_n^0$ (i.e. on β_n^0 and θ_n^0), as well as on $g_n(A \mid W)$, an estimator of the nuisance parameter $g_0(A \mid W) \equiv P_0(A \mid W)$. Therefore, we have

$$\begin{aligned} \log\bar{Q}_{\beta_n^0,\theta_n^0}(\epsilon)(A,W) &= (\beta_n^0+\epsilon)A + \log\theta_n^0(W) + \epsilon r^*_{\bar{Q}_n^0,g_n}(W) \\ &= \beta_n^0 A + \log\theta_n^0(W) + \epsilon(A + r^*_{\bar{Q}_n^0,g_n}(W)). \end{aligned} \tag{5.2}$$

Therefore our clever covariate is given by $A + r^*_{\bar{Q}_n^0,g_n}(W)$. To construct $r^*_{\bar{Q}_n^0,g_n}(W)$, we require that when $\epsilon = 0$, we have the initial density, $\bar{Q}_n^0(A,W)$ and that the score of this submodel with parameter ϵ has score at $\epsilon = 0$ equal to the efficient score of the target parameter β_0. Because we have not yet made any assumptions about the distribution of Y, so far, this applies to both types of outcomes (binary and count) and corresponding parameters of interest.

We fit ϵ with MLE, which results in a first step update of Q_n^0, which plays the role of the initial estimator in the next update step. As described in the previous chapter, in the TMLE algorithm, we iteratively update an initial estimator until the next MLE of ϵ is close

to zero. Therefore, for each of $k = 0 \dots K$ iterations, we compute an updated $\bar{Q}_{\beta_n^k, \theta_n^k}(\epsilon_n^k)$, using the function $r^*_{\bar{Q}_n^k, g_n}$. At convergence, the final updated estimator of $\bar{Q}_0$ (for either outcome type/parameter) is given by $\bar{Q}_n^* \equiv \bar{Q}_n^K$ and we evaluate β_n^* by substitution.

5.3 Conditional Relative Risk Parameters in a Semi-parametric Regression Model

In this section, we assume Y is binary and follows a binomial distribution and that the logarithm of its expected value can be modeled by a linear combination of exposure and covariates (i.e. we use a log-binomial regression model). We focus on the case in which $m_{\beta_0}(A, V) = \beta_0 A$. Therefore, the parameter of interest, β_0, in our semi-parametric multiplicative regression model is the change in log conditional relative risk associated with a one unit change in exposure A. Below, in Section 5.3.1, we first construct the efficient score and efficient influence curve for β_0. Then in Section 5.3.2, we derive the TMLE for this parameter. Therefore, we present the form of the parametric fluctuation submodel through the initial estimator, find the score of this fluctuation submodel and derive the clever covariate such that the score of the fluctuation submodel equals the efficient score. This clever covariate allows us to define the specific fluctuation submodel for the TMLE step.

5.3.1 Constructing the Efficient Score and Efficient Influence Curve

Recall that the probability distributions of Y, given A, W, in the semi-parametric model are indexed by a finite dimensional parameter β and infinite dimensional parameter θ. Note that for semi-parametric models, the efficient influence curve, $D^*_{\bar{Q}_0, g_0}(O)$ is defined as

$$D^*_{\bar{Q}_0, g_0}(O) = -\left[\frac{d}{d\beta_0} E(S^*_{\bar{Q}_0, g_0}(O))\right]^{-1} S^*_{\bar{Q}_0, g_0}(O),$$

where $S^*_{\bar{Q}_0, g_0}(O)$ denotes the efficient score given by

$$S^*_{\bar{Q}_0, g_0}(O) = S_{\beta_0} - \Pi(S_{\beta_0} \mid T_{nuis}). \tag{5.3}$$

Here $S_{\beta_0} \equiv S_{\beta_0}(Y \mid A, W) = \frac{d}{d\beta_0} \log Q_{\beta_0, \theta_0}(Y \mid A, W)$ is the score of the parameter of interest, β_0, and T_{nuis} is the nuisance tangent space, viewed as a subspace of the Hilbert space $L^2_0(P_0)$ endowed with the inner product $\langle h_1, h_2 \rangle = E_0 h_1 h_2(O)$. Because the data generating distribution is indexed by a parameter of interest β_0 and a variation independent nuisance parameter, the efficient score is equal to the score of the parameter of interest minus the projection of this score onto the nuisance tangent space. Recall that a projection of a function S on a subspace T_{nuis} of a Hilbert space is uniquely defined by 1) the projection

that is an element of the subspace T_{nuis}, and 2) $S - \Pi(S \mid T_{nuis}) \perp T_{nuis}$. And note that T_{nuis} is the direct sum of the three orthogonal spaces involving each of the nuisance parameters:

$$T_{nuis} = T_W \bigoplus T_{A|W} \bigoplus T_\theta,$$

where T_W consists of all functions in $L_0^2(P_0)$ of W with mean zero; $T_{A|W}$ consists of all functions in $L_0^2(P_0)$ of (A, W) with conditional mean zero, given W; and T_θ is the tangent space spanned by all the scores of parametric submodels through P_0 that only fluctuate θ_0.

Therefore, we can write Equation 5.3 as:

$$S^*_{\bar{Q}_0, g_0}(O) = S_{\beta_0} - \left[\prod(S_{\beta_0} \mid T_W) + \prod(S_{\beta_0} \mid T_{A|W}) + \prod(S_{\beta_0} \mid T_\theta)\right]. \tag{5.4}$$

Given Equation (5.4), we carry out the following steps to construct the efficient score.

1. **First, we calculate S_{β_0}.** The probability distribution $P_{\beta,\theta}(Y \mid A, W)$ is indexed by β_0 and function θ_0. We have $\log P_{\beta_0,\theta_0}(Y = 1 \mid A, W) = \log \theta_0(W) + \beta_0 A$. It follows that $\frac{d}{d\beta_0} \log P_{\beta_0,\theta_0}(Y \mid A, W) = \frac{A}{1-\bar{Q}_{\beta_0,\theta_0}}(Y - \bar{Q}_{\beta_0,\theta_0}(A, W))$.

2. **Next, we calculate the nuisance scores for each nuisance parameter.**

   We do this by fluctuating each of the nuisance parameters. First, to calculate T_W, we note that the probability distribution $P_{W,0}$ varies over a non-parametric model. Therefore, we can fluctuate it as follows:

   $$P_0(\epsilon)(W) = (1 + \epsilon h_1(W)) P_0(W),$$

   where $h_1(W)$ is any function of W such that $E_0(h_1(W)) = 0$ and $E_0(h_1^2(W)) < \infty$. Then the nuisance score generated by this parametric submodel is given by h_1. This shows that $T_W = \{W \to h_1(W) : E_0 h_1(W) = 0, E_0 h_1^2(W) < \infty\}$.

   To calculate $T_{A|W}$, we note that $P_{0,A|W}$ also varies over a non-parametric model, so that we can fluctuate it as follows:

   $$P_0(\epsilon)(A \mid W) = (1 + \epsilon h_2(A, W)) P_0(A \mid W),$$

   where $h_2(A, W)$ is any function of A and W such that $E_0(h_2(A, W) \mid W) = 0$ and $E_0 h_2^2(A, W) < \infty$. As above, it follows that $T_{A|W} = \{(A, W) \to h_2(A, W) : E_0(h_2 \mid W) = 0, E_0 h_2^2 < \infty\}$.

Finally, to calculate T_θ, we consider submodels $\bar{Q}_0(\epsilon)(Y \mid A, W)$ implied by $\log \bar{Q}_0(\epsilon)(A, W) = \log \theta_0(W) + \beta_0 A + \epsilon h_3(W)$ for an arbitrary function h_3. Notice that this implies a submodel in our semi-parametric regression model. It is straightforward to show that the score of this submodel at $\epsilon = 0$ equals $1/(1 - \bar{Q}_0(A, W))h_3(W)(Y - \bar{Q}_0(A, W))$. This shows that $T_\theta = \{1/(1 - \bar{Q}_0(A, W))h_3(W)(Y - \bar{Q}_0(A, W)) : h_3\}$.

Therefore, we can conclude that the nuisance tangent space is defined as:

$$\begin{aligned} T_{nuis} &= \{h_1(W) : E(h_1(W)) = 0; E(h_1^2(W)) < \infty\} \\ &+ \{h_2(A, W) : E(h_2(A, W) \mid W) = 0; E(h_2^2(A, W)) < \infty\} \\ &+ \left\{ \frac{h_3(W)}{1 - \bar{Q}_0}(Y - \bar{Q}_0(A, W)) : E(h_3(W)) = 0; E(h_3^2(W)) < \infty \right\}. \end{aligned}$$

3. **Finally, we calculate the projections of S_{β_0} onto each of the nuisance tangent spaces, as seen in (5.4).**

The first two are straight-forward. Because S_{β_0} has conditional mean, given A, W, equal to zero, it follows that it is orthogonal to T_W and $T_{A|W}$. Therefore we have

$$\begin{aligned} \prod(S_{\beta_0} \mid T_W) &= 0 \\ \prod(S_{\beta_0} \mid T_{A|W}) &= 0. \end{aligned}$$

For $\prod(S_{\beta_0} \mid T_\theta)$ the calculation is more complicated. We define $V \equiv V(Y, A, W)$ as a function of the data such that $E_0(V \mid A, W) = 0$. As repeatedly used and shown in van der Laan and Robins [2003]: any function $S(B, Pa(B))$ of a binary variable B and other variables $Pa(B)$, which has conditional mean zero, given $Pa(B)$, can be written as $(S(1, Pa(B)) - S(0, Pa(B))(B - P(B = 1 \mid Pa(B)))$. Let $h_V(A, W) = (V(1, A, W) - V(0, A, W))$ so that $V - E_0(V|A, W) = h_V(A, W)(Y - \bar{Q}_0)$. Thus,

$$\begin{aligned} \prod(V \mid T_\theta) &= \prod(V - E_0(V|A, W)|T_\theta) \\ &= \prod(\{V(1, A, W) - V(0, A, W)\}(Y - \bar{Q}_0)|T_\theta). \end{aligned}$$

We have

$$\prod(V - E_0(V|A, W)|T_\theta) = \prod\left(h_V(Y - \bar{Q}_0)|\left\{\frac{(Y - \bar{Q}_0)h_3(W)}{1 - \bar{Q}_0} : h_3(W)\right\}\right). \quad (5.5)$$

In particular, if $V = S_{\beta_0,\theta_0}$, we have $V = A/(1-\bar{Q}_0)(Y-\bar{Q}_0)$ so that $h = A/(1-\bar{Q}_0)$.

Thus, we need to find the function h_3^* such that for all h_3

$$\left\langle h_V(A,W)(Y-\bar{Q}_0) - (Y-\bar{Q}_0)h_3^*(W), (Y-\bar{Q}_0)h_3(W)\right\rangle = 0.$$

Therefore, we want to find h_3^* such that

$$\begin{aligned}
E_0\left[\left\{h(A,W)(Y-\bar{Q}_0) - \frac{h_3^*(W)}{1-\bar{Q}_0}(Y-\bar{Q}_0)\right\}\frac{h_3(W)}{1-\bar{Q}_0}(Y-\bar{Q}_0)\right] &= 0 \text{ for all } h_3(W)\\
E_0\left[\left(h(A,W) - \frac{h_3^*(W)}{1-\bar{Q}_0}\right)\frac{h_3(W)}{1-\bar{Q}_0}(Y-\bar{Q}_0)^2\right] &= 0 \text{ for all } h_3(W)\\
E_0\left[\left(h(A,W) - \frac{h_3^*(W)}{1-\bar{Q}_0}\right)\frac{h_3(W)}{1-\bar{Q}_0}\sigma^2(A,W)\right] &= 0 \text{ for all } h_3(W)\\
E_0\left[\left(h(A,W)\frac{\sigma^2}{1-\bar{Q}_0} - \frac{h_3^*(W)}{(1-\bar{Q}_0)^2}\sigma^2\right)h_3(W)\right] &= 0 \text{ for all } h_3(W)\\
E_0\left[E\left[\frac{h(A,W)}{1-\bar{Q}_0}\sigma^2|W\right] - h_3^*(W)E_0\left[\frac{\sigma^2}{(1-\bar{Q}_0)^2}|W\right]h_3(W)\right] &= 0 \text{ for all } h_3(W),
\end{aligned}$$

where $\sigma^2(A,W) = \mathrm{VAR}_0(Y \mid A,W)$. Therefore:

$$h_3^*(W) = \frac{E_0\left(\frac{h(A,W)\sigma^2}{1-\bar{Q}_0}|W\right)}{E_0\left(\frac{\sigma^2}{(1-\bar{Q}_0)^2}|W\right)}.$$

Plugging into (5.5), we see that

$$\begin{aligned}
\prod(V - E_0(V|A,W)|T_\theta) &= \frac{E_0\left[h_V(A,W)\sigma^2/(1-\bar{Q}_0)|W\right]}{E\left[\sigma^2/(1-\bar{Q}_0)^2|W\right]}\frac{(Y-\bar{Q}_0)}{1-\bar{Q}_0}\\
&= \frac{E_0\left[(V(1,A,W)-V(0,A,W))\frac{\sigma^2}{1-\bar{Q}_0}|W\right]}{E_0\left[\frac{\sigma^2}{(1-\bar{Q}_0)^2}|W\right]}\frac{(Y-\bar{Q}_0)}{1-\bar{Q}_0}\\
&= \frac{E_0\left[(V(1,A,W)-V(0,A,W))\bar{Q}_0|W\right]}{E_0\left[\frac{\bar{Q}_0}{(1-\bar{Q}_0)}|W\right]}\frac{(Y-\bar{Q}_0)}{1-\bar{Q}_0}.
\end{aligned}$$

So this tells us how to project any V onto the nuisance tangent space T_{nuis}. Now we know how to project S_{β_0} on T_θ, so we have

$$\begin{aligned}\prod(S_{\beta_0}|T_\theta) &= \frac{E_0\left[\frac{A\sigma^2}{(1-\bar{Q}_0)^2}|W\right]}{E_0\left[\frac{\sigma^2}{(1-\bar{Q}_0)^2}|W\right]}\frac{(Y-\bar{Q}_0)}{1-\bar{Q}_0} \\ &= \frac{E_0\left[\frac{A\bar{Q}_0}{(1-\bar{Q}_0)}|W\right]}{E_0\left[\frac{\bar{Q}_0}{1-\bar{Q}_0}|W\right]}\frac{(Y-\bar{Q}_0)}{1-\bar{Q}_0}.\end{aligned}$$

Now finally, by plugging $S_{\beta_0} = A/(1-\bar{Q}_0)(Y-\bar{Q}_0)$ and all projections into (5.4), we can write down the efficient score:

$$\begin{aligned}S^*_{\bar{Q}_0,g_0}(O) &= S_{\beta_0} - \prod(S_{\beta_0}|T_{nuis}) \text{ as above} \\ &= S_{\beta_0} - \left[\prod(S_{\beta_0}|T_W) - \prod(S_{\beta_0}|T_{A|W}) - \prod(S_{\beta_0}|T_\theta)\right] \\ &= \frac{1}{1-\bar{Q}_0}\left(A - \frac{E_0\left[\frac{A\bar{Q}_0}{(1-\bar{Q}_0)}|W\right]}{E_0\left[\frac{\bar{Q}_0}{1-\bar{Q}_0}|W\right]}\right)(Y-\bar{Q}_0).\end{aligned}$$

As noted in the last chapter, it is of interest to note that this can also be represented as

$$h^*(A \mid W)(Y\bar{Q}_0(0,W)/\bar{Q}_0(A,W) - \bar{Q}_0(0,W)),$$

where

$$h^*(A \mid W) = \frac{\bar{Q}_0}{\bar{Q}_0(0,W)(1-\bar{Q}_0)}\left(A - \frac{E_0\left[\frac{A\bar{Q}_0}{(1-\bar{Q}_0)}|W\right]}{E_0\left[\frac{\bar{Q}_0}{1-\bar{Q}_0}|W\right]}\right)$$

is a function satisfying $E_0(h^*(A \mid W) \mid W) = 0$. This representation can be used to prove that $S^*_{\beta_0,\theta,g_0}(O)$ has mean zero for all θ, thereby proving the double robustness of the efficient score as estimating function for β_0.

Recall, this entire derivation assumes that $m_{\beta_0}(A,V) = \beta_0 A$. For the more general case, if $\log\bar{Q}_0 = m_{\beta_0} + \log\theta_0$, the efficient score is given by

$$S^*_{\bar{Q}_0,g_0}(O) = \frac{1}{1-\bar{Q}_0}\left(\frac{d}{d\beta_0}m_{\beta_0} - \frac{E_0\left[\frac{\frac{d}{d\beta_0}m_{\beta_0}\bar{Q}_0}{(1-\bar{Q}_0)}|W\right]}{E_0\left[\frac{\bar{Q}_0}{1-\bar{Q}_0}|W\right]}\right)(Y-\bar{Q}_0).$$

Recall the efficient influence curve is then given by

$$D^*_{\bar{Q}_0,g_0}(O) = -E\left[\frac{d}{d\beta_0}S^*_{\bar{Q}_0,g_0}(O)\right]^{-1}S^*_{\bar{Q}_0,g_0}(O).$$

5.3.2 Deriving the TMLE

Defining the parametric fluctuation submodel

In this case, we assume a binomial regression model for the parametric fluctuation submodel given in (5.2). Therefore, we have

$$P_{\beta_n^0,\theta_n^0}(\epsilon)(Y \mid A,W) = \left[e^{(\beta_n^0+\epsilon)A}\theta_n^0(\epsilon)(0,W)\right]^Y \left[1 - e^{(\beta_n^0+\epsilon)A}\theta_n^0(\epsilon)(0,W)\right]^{1-Y},$$

where, as defined above, $\theta_n^0(\epsilon)(0,W) = \theta_n^0(0,W)e^{\epsilon r^*_{\bar{Q}_n^0,g_n}(W)}$.

Of course, we can also write down the parametric fluctuation model at the truth, which is given by

$$P_{\beta_0,\theta_0}(\epsilon)(Y \mid A,W) = \left[e^{(\beta_0+\epsilon)A}\theta(\epsilon)(0,W)\right]^Y \left[1 - e^{(\beta_0+\epsilon)A}\theta(\epsilon)(0,W)\right]^{1-Y}, \qquad (5.6)$$

where now $\theta_0(\epsilon)(0,W) = \theta_0(0,W)e^{\epsilon r^*_{\bar{Q}_0,g_0}(W)}$. We wish to determine the function $r^*_{\bar{Q}_0,g_0}(W)$ so that the score at $\epsilon = 0$ equals the efficient influence curve at P_0, and we do this below.

Calculating the score of the parametric fluctuation submodel at $\epsilon = 0$

To calculate the score of the fluctuation model in Equation 5.6, we take logs, take the derivative with respect to ϵ and evaluate at $\epsilon = 0$.

At $Y = 1$, we have

$$\begin{aligned}
P_{\beta_0,\theta_0}(\epsilon)(Y=1|A,W) &= e^{(\beta_0+\epsilon)A}\theta_0(0,W)e^{\epsilon r^*_{\bar{Q}_0,g_0}(W)} \\
\log P_{\beta_0,\theta_0}(\epsilon)(Y=1|A,W) &= (\beta_0+\epsilon)A + \log\theta_0(0,W) + \epsilon r^*_{\bar{Q}_0,g_0}(W) \\
\frac{d}{d\epsilon}\log P_{\beta_0,\theta_0}(\epsilon)(Y=1|A,W)|_{\epsilon=0} &= A + r^*_{\bar{Q}_0,g_0}(W).
\end{aligned}$$

At $Y = 0$, we have

$$\begin{aligned}
P_{\beta_0,\theta_0}(\epsilon)(Y=0|A,W) &= 1 - P_{\beta_0,\theta_0}(\epsilon)(Y=1|A,W) \\
&= 1 - \left[e^{(\beta_0+\epsilon)A}\theta_0(0,W)e^{\epsilon r^*_{\bar{Q}_0,g_0}(W)}\right] \\
\frac{d}{d\epsilon}\log P_{\beta_0,\theta_0}(\epsilon)(Y=0|A,W)|_{\epsilon=0} &= \frac{-\bar{Q}_{\beta_0,\theta_0}}{1-\bar{Q}_{\beta_0,\theta_0}}\left(A + r^*_{\bar{Q}_0,g_0}(W)\right).
\end{aligned}$$

Therefore, the score of the fluctuation model at $\epsilon = 0$ is given by

$$\begin{aligned}\frac{d}{d\epsilon}\log\bar{Q}_{\beta_0,\theta_0}(\epsilon)(Y \mid A,W) &= Y\left[A + r^*_{\bar{Q}_0,g_0}(W)\right] - (1-Y)\left[\frac{-\bar{Q}_{\beta_0,\theta_0}}{1-\bar{Q}_{\beta_0,\theta_0}}(A + r^*_{\bar{Q}_0,g_0}(W))\right] \\ &= \frac{1}{1-\bar{Q}_{\beta_0,\theta_0}}\left[A + r^*_{\bar{Q}_0,g_0}(W)\right][Y - \bar{Q}_{\beta_0,\theta_0}].\end{aligned}$$

Next, we want to determine the function $r^*_{Q_0,g_0}$ that makes this score equal the efficient score. We do this below.

Constructing the parametric fluctuation submodel with score spanning the efficient score

If we assume $\bar{Q}_0 = \exp(m_{\beta_0}(A,W))\theta_0(W)$, and we use as submodel $\log\bar{Q}_0(\epsilon) = m_{\beta_0+\epsilon} + \log\theta_0 + \epsilon r$, then the score equals $(d/d\beta_0 m_{\beta_0} + r)(Y - \bar{Q}_0)/(1-\bar{Q}_0)$. Thus, to arrange that this score equals the efficient score, we have

$$r^*_{\bar{Q}_0,g_0} = -\frac{E_0\left[\frac{d/d\beta_0 m_{\beta_0}\bar{Q}_0}{(1-\bar{Q}_0)}|W\right]}{E_0\left[\frac{\bar{Q}_0}{1-\bar{Q}_0}|W\right]}.$$

We can then compute the clever covariate, which is given by

$$H^*_{\bar{Q}_0,g_0} = A - \frac{E_0\left[\frac{A\bar{Q}_0}{(1-\bar{Q}_0)}|W\right]}{E_0\left[\frac{\bar{Q}_0}{1-\bar{Q}_0}|W\right]}.$$

Also, in the general case, for any $m_{\beta_0}(A,V)$, the clever covariate $H^*_{\bar{Q}_0,g_0} \equiv H^*_{\beta_0,\theta_0,g_0}$ is given by

$$H^*_{\beta_0,\theta_0,g_0} = \frac{d}{d\beta}m_{\beta_0}(A,V) - \frac{E_0\left[\frac{\bar{Q}_0(A,W)}{1-\bar{Q}_0(A,W)}\frac{d}{d\beta}m_{\beta_0}(A,V)|W\right]}{E_0\left[\frac{\bar{Q}_0(A,W)}{1-\bar{Q}_0(A,W)}|W\right]}.$$

We add the clever covariate to $\log\bar{Q}^k_n$, which we hold constant and then estimate the fluctuation parameter ϵ by maximum likelihood. This yields the $k^{th}+1$ update of the initial fit. We then repeat the fluctuation until ϵ is very close to zero, at which point we have our final targeted estimate of the density, $\bar{Q}^*_n(A,W)$. The TMLE, β^*_n is then evaluated by substitution, or to calculate, we can simply add the sum of all ϵ estimates to β^0_n. We can then estimate the variance of the influence curve with the empirical variance and calculate standard errors, p-values and confidence intervals for β^*_n.

5.4 Conditional Incidence Rate Parameters in a Semi-parametric Regression Model

In this section, we assume Y is a count of events and that Y has a Poisson distribution. We again assume that the logarithm of its expected value can be modeled by a linear combination of exposure and covariates, but now we are assuming this is a Poisson regression model. Our parameter, β_0, in this case (again focusing on the case in which $m_{\beta_0}(A,V) = \beta_0 A$), is the log conditional incidence rate associated with a unit change in exposure. Below, in Section 5.4.1, we construct the efficient score and efficient influence curve for this β_0. Then in Section 5.4.2, we derive the TMLE for β_0 in the semi-parametric Poisson regression model. As discussed in the introduction, we can apply the resulting TMLE for count data to binary data, which is illustrated in Chapter 3. In this case, the TMLE is no longer efficient, but it is a valid estimator of conditional relative risk in a semi-parametric regression model. In Section 5.4.3 we discuss the implications in more detail.

5.4.1 Constructing the Efficient Score and Efficient Influence Curve

As above, the efficient influence curve, $D^*_{\bar{Q}_0,g_0}(O)$ is defined as

$$D^*_{\bar{Q}_0,g_0}(O) = -\left[\frac{d}{d\beta_0}E_0(S^*_{\bar{Q}_0,g_0}(O))\right]^{-1} S^*_{\bar{Q}_0,g_0}(O).$$

where $S^*_{\bar{Q}_0,g_0}(O)$ is the efficient score. In order to construct $S^*_{\bar{Q}_0,g_0}(O)$ when we assume that Y is a count of events and follows a Poisson distribution, we need that the score of the Poisson-distribution fluctuation at $\epsilon = 0$ is an element of the orthogonal complement of the nuisance tangent space. Therefore, we carry out the following steps:

1. **First, we find the set of all scores of the Poisson fluctuation at $\epsilon = 0$.**

   For the TMLE in the Poisson semi-parametric regression model, we need a parametric fluctuation model through the initial estimator that is a submodel of the semi-parametric Poisson regression model. We select submodels of the form:

$$\begin{aligned} P_{\beta_n^0,\theta_n^0}(\epsilon)(Y \mid A,W) &= \frac{[\bar{Q}_{\beta_n^0,\theta_n^0}(\epsilon)(A,W)]^Y}{Y!}\exp(-(\bar{Q}^0_{n_{\beta_n^0,\theta_n^0}}(\epsilon)(A,W)) \\ &= \frac{[e^{(\beta_n^0+\epsilon)A}\theta_n^0(\epsilon)(W)]^Y}{Y!}\exp(-(e^{(\beta_n^0+\epsilon)A}\theta_n^0(\epsilon)(W)), \end{aligned}$$

   where $\theta_n^0(\epsilon)(W) = \theta_n^0(W)\exp(\epsilon r^*_{\bar{Q}_n^0,g_n}(W))$. Or, at the truth, we have

$$P_{\beta_0,\theta_0}(\epsilon)(Y \mid A,W) = \frac{[e^{(\beta_0+\epsilon)A}\theta_0(\epsilon)(W)]^Y}{Y!}\exp(-(e^{(\beta_0+\epsilon)A}\theta_0(\epsilon)(W)), \qquad (5.7)$$

where $\theta^0(\epsilon)(W) = \theta_0(W)\exp(\epsilon r^*_{\bar{Q}_0,g_0}(W))$. To calculate the score of the fluctuation model in Equation 5.7, we first take logs, so that we have

$$\log P_{\beta_0,\theta_0}(\epsilon)(Y \mid A,W) = Y\left((\beta_0+\epsilon)A + log\theta_0 + \epsilon r^*_{\bar{Q}_0,g_0}\right) - e^{(\beta_0+\epsilon)A}\theta_0 e^{\epsilon r^*_{\bar{Q}_0,g_0}} - \log Y!.$$

Therefore, to obtain the score at $\epsilon = 0$, we have

$$\begin{aligned}
\frac{d}{d\epsilon} log P_{\beta_0,\theta_0}(\epsilon)(Y \mid A,W) &= YA + r^*_{\bar{Q}_0,g_0} - \left(e^{(\beta_0+\epsilon)A}A\theta_0 e^{\epsilon r^*_{\bar{Q}_0,g_0}} + e^{(\beta_0+\epsilon)A}\theta_0 e^{\epsilon r^*_{\bar{Q}_0,g_0}} r^*_{\bar{Q}_0,g_0}\right) \\
&= Y(A + r^*_{\bar{Q}_0,g_0}) - (e^{(\beta_0+\epsilon)A}\theta_0 e^{\epsilon r^*_{\bar{Q}_0,g_0}})(A + r^*_{\bar{Q}_0,g_0}) \\
&= (Y - (e^{(\beta_0+\epsilon)A}\theta_0 e^{\epsilon r^*_{\bar{Q}_0,g_0}}))(A + r^*_{\bar{Q}_0,g_0}). \\
\frac{d}{d\epsilon} log P_{\beta_0,\theta_0}(\epsilon)(Y \mid A,W)|_{\epsilon=0} &= (Y - e^{\beta_0 A}\theta_0)(A + r^*_{\bar{Q}_0,g_0}) \\
&= (Y - \bar{Q}_{\beta_0,\theta_0})(A + r^*_{\bar{Q}_0,g_0}).
\end{aligned}$$

Therefore, the set of all scores at $\epsilon = 0$ for a Poisson fluctuation, which is indexed by a function $r^*_{\bar{Q}_0,g_0}(W)$, can be written as follows when $m_{\beta_0}(A,V) = \beta_0 A$:

$$\frac{d}{d\epsilon}\log P_0(\epsilon)|_{\epsilon=0} = \{A + r^*_{\bar{Q}_0,g_0}\}(Y - \bar{Q}_0),$$

or in the general case as:

$$\frac{d}{d\epsilon}\log P_0(\epsilon)|_{\epsilon=0} = \{\frac{d}{d\beta_0}m_{\beta_0} + r^*_{\bar{Q}_0,g_0}\}(Y - \bar{Q}_0).$$

2. **Next, we need the orthogonal complement of the nuisance tangent space.**

   From a result in van der Laan et al. [2004b], for any $m_{\beta_0}(A,V)$, the orthogonal complement of the nuisance tangent space consists of functions given by $m^*_{\beta_0}h(A \mid W)(Y - \bar{Q}_0)$, indexed by functions $h(A \mid W)$ with conditional mean zero given W, where $m^*_{\beta_0} = e^{-m_{\beta_0}}$.

3. **Therefore, we need to find $h(A \mid W)$ with $E(h(A \mid W) \mid W) = 0$ such that**

$$m^*_{\beta_0} h(A \mid W)(Y - \bar{Q}_0) = \{\frac{d}{d\beta_0} m_{\beta_0} + r^*_{\bar{Q}_0, g_0}\}(Y - \bar{Q}_0).$$

Therefore, we have

$$h(A \mid W) = \frac{\frac{d}{d\beta_0} m_{\beta_0} + r^*_{\bar{Q}_0, g_0}}{m^*_{\beta_0}}. \tag{5.8}$$

Then,

$$0 = E\left[h(A \mid W) \mid W\right] = E\left[\frac{\frac{d}{d\beta_0} m_{\beta_0} + r^*_{\bar{Q}_0, g_0}}{m^*_{\beta_0}} \mid W\right]. \tag{5.9}$$

4. **Finally, we find $r^*_{\bar{Q}_0, g_0}(W)$.**

The only $r^*_{\bar{Q}_0, g_0}(W) \equiv r^*_{\beta_0, \theta_0, g_0}(W)$ that makes (5.9) true is given by

$$r^*_{\beta_0, \theta_0, g_0}(W) \quad = \quad -\frac{E_{g_0}\left(\frac{1}{m^*_{\beta_0}} \frac{d}{d\beta_0} m_{\beta_0} \mid W\right)}{E_g\left(\frac{1}{m^*_{\beta_0}} \mid W\right)}.$$

Plugging this into (5.8), we see that this $r^*_{\beta_0, \theta_0, g_0}(W)$ corresponds with a function $h(A \mid W)$, which we refer to as $h_{opt}(\beta_0, \theta_0, g_0)(W)$, given by

$$h_{opt}(\beta_0, \theta_0, g_0)(W) = \frac{1}{m^*_{\beta_0}} \left\{\frac{d}{d\beta_0} m_{\beta_0} + r^*_{\beta_0, \theta_0, g_0}(W)\right\},$$

which indeed has conditional mean zero, given W. Therefore, we can write down the efficient score as given by

$$S^*_{\beta_0, \theta_0, g_0}(O) = \left(\frac{d}{d\beta_0} m_{\beta_0} - \frac{E\left[e^{m_{\beta_0}} \frac{d}{d\beta_0} m_{\beta_0} | W\right]}{E\left[e^{m_{\beta_0}} | W\right]}\right)(Y - \bar{Q}_{\beta_0, \theta_0}).$$

When we assume $m_{\beta_0}(A, V) = \beta_0 A$, it is given by

$$S^*_{\beta_0, \theta_0, g_0}(O) = \left(A - \frac{E\left[e^{\beta_0 A} A | W\right]}{E\left[e^{\beta_0 A} | W\right]}\right)(Y - \bar{Q}_{\beta_0, \theta_0}).$$

5.4.2 Deriving the TMLE

When $m_{\beta_0}(A,V) = \beta_0 A$, the clever covariate, which is given by $A + r^*_{\beta_0,\theta_0,g_0}$ is

$$H^*_{\beta_0,\theta_0,g_0} = A - \frac{E\left[e^{\beta_0 A}A|W\right]}{E\left[e^{\beta_0 A}|W\right]}.$$

In the general case, for any assumed $m_{\beta_0}(A,V)$, the clever covariate is given by $\frac{d}{d\beta_0}m_{\beta_0}(A,V)+ r^*_{\beta_0,\theta_0,g_0}$, so the clever covariate is given by

$$H^*_{\beta_0,\theta_0,g_0} = \frac{d}{d\beta_0}m_{\beta_0}(A,V) - \frac{E\left[e^{m_{\beta_0}(A,V)}\frac{d}{d\beta_0}m_{\beta_0}(A,V)|W\right]}{E\left[e^{m_{\beta_0}(A,V)}|W\right]}.$$

For each of k iterations of updating our initial estimate of $\bar{Q}_0(A,W)$, we estimate the clever covariate, so that when $m_{\beta_0}(A,V) = \beta_0 A$, we have:

$$H^*_{\beta^k_n,\theta^k_n,g_n} = A - \frac{E\left[e^{\beta^k_n A}A|W\right]}{E\left[e^{\beta^k_n A}|W\right]}.$$

As described above for the "correct" TMLE, we add this clever covariate to $\log\bar{Q}^k_n$, which we hold constant and then estimate the fluctuation parameter ϵ by maximum likelihood. This yields the $k^{th}+1$ update of the initial fit. We then repeat the fluctuation until ϵ is very close to zero, at which point we have our final targeted estimate of the density, $\bar{Q}^*_n(A,W)$. The TMLE, β^*_n is then evaluated by substitution, or to calculate, we can simply add the sum of all ϵ estimates to β^0_n. We can then estimate the variance of the influence curve with the empirical variance and calculate standard errors, p-values and confidence intervals for β^*_n.

5.4.3 Remarks on Applying the TMLE of the Semi-parametric Poisson Regression to a Binary Outcome

As discussed above, as well as in Chapter 3, we can implement the TMLE described in this section - the TMLE that assumes Y is a count of events - to estimate our parameter of interest when Y is actually binary, the change in log conditional relative risk associated with a unit change in exposure. When we apply this TMLE to a binary outcome, the Poisson model is always wrong. However, the TMLE is still a DR and asymptotically linear estimator of the relative risk parameter of interest. It is not efficient, however. We make this trade-off - of using the TMLE for a count outcome rather than the TMLE for a binary outcome - because the Poisson-based TMLE is much more practical in application. It is much more computationally stable as the log-binomial model often suffers from non-convergence problems.

However, for both TMLE's, the efficient score equations are derived from the overall efficient score equation, which makes no distributional assumptions on the distributional form of the residuals. For a proof of this property, see the appendix in Tuglus et al. [2011]. Consequently, because inference for the TMLE is based on solving the efficient score equation, the inference remains valid, regardless of the form of the parametric submodel.

5.5 Discussion

This chapter has provided theoretical details for defining and implementing two TMLE's for estimating parameters relating conditional relative risk to exposure defined by a semi-parametric multiplicative regression model. We began by assuming a binary outcome, Y, for which $P_0(Y = 1|A, W)$ is modeled as a semi-parametric multiplicative model, such that $P_0(Y = 1|A, W) \equiv \bar{Q}_0(A, W) = e^{m_{\beta_0}(A,V)}\theta_0(W)$. For the first TMLE, we correctly assumed that Y has a binomial distribution, and for the second TMLE we incorrectly assumed that Y has a Poisson distribution, as if it were actually a count of events. When we consider $m_{\beta_0}(A, V) = \beta_0 A$, the latter of the two TMLE's actually is an estimator for the change in the log conditional incidence rate associated with a unit change in A, rather than estimator or our parameter of interest, the change in the log conditional relative risk associated with a unit change in A. However, we can apply this second, count-data, or "practical" TMLE to our parameter of interest for binary data, and we can show that it remains a DR and asymptotically linear estimator. See Tuglus et al. [2011] for an illustration and proof.

For both TMLE's, the distinguishing component is the parametric submodel through the initial estimator of the density of the data that has score equal to the efficient score. Therefore, in this chapter, for each of the TMLE's, we have provided theoretical details for constructing the efficient score for the corresponding true parameter of interest. We have also derived in each case the corresponding clever covariate for carrying out the fluctuation.

Chapter 6

Conclusion

As a whole, this dissertation has demonstrated the importance of the estimator selection when estimating parameters under a lack of positivity. It began with a thorough investigation of the positivity assumption, how different estimators respond to positivity violations, a diagnostic for bias due to positivity violations, based on the parametric bootstrap, and a discussion of options for responding to any remaining, diagnosed bias due positivity violations. This dissertation also delved more deeply into the relative performance of TMLE's under lack of positivity, focusing on three common parameters of interest. It described the theoretical features of the TMLE's which indicate why they tend to be relatively robust under positivity violations and illustrated the robustness with a wide variety of simulations. Many of the simulations were based on studies presented in the literature. This allowed us to benchmark our performance against existing work and avoid the criticism of developing simulations designed to illustrate our points. We also tweaked the existing simulations to make the estimation problem even more challenging, providing an even more valuable demonstration of the relative performance of TMLE's in a variety of settings. Finally, this dissertation also delved into the theoretical details on which TMLE methodology is based. Focusing on two different TMLE's of conditional relative risk in a semi-parametric multiplicative regression model, the final chapter provided an in-depth look at how to construct important TMLE features - the efficient score, efficient influence curve and the clever covariate that defines the targeted fluctuation. Overall, the following summarizes the main conclusions that can be drawn from the work presented across all chapters in this dissertation:

- The estimation method can really matter, particularly in observational studies.
- Positivity violations are a common challenge in observational data and can threaten valid inference for many parameters of interest.
- Bias due to positivity violations often goes undiagnosed. However, the parametric bootstrap is a valuable tool that can identify bias not necessarily evident with other diagnostic approaches.

- Different estimators are affected differently by lack of positivity (and by bounding g_n).
- TMLE's and C-TMLE's are more robust to violations of the positivity assumption.
- C-TMLE's provide an innovative "black-box" approach for estimating the censoring mechanism, preferring covariates that are associated with Y and A.
- TMLE's, as well as other estimators, can be combined with data-adaptive methods such as super learning, which improves robustness due to model misspecification.
- The parametric bootstrap diagnostic can be a valuable tool for evaluating different estimators and helping to select among them.
- When bias in a chosen estimator is still diagnosed, analysts must consider alternative parameters with better identifiability (e.g. modify adjustment set or sample). These parameters may be more appropriate to research goals.

Bibliography

A.J. Barros and V.N. Hirakata. Alternatives for logistic regression in cross-sectional studies: an empirical comparison of models that directly estimate the prevalence ratio. *BMC Med Res Methodol*, 3:21, Oct 2003. doi: 10.1186/1471-2288-3-21.

O. Bembom and M.J. van der Laan. A practical illustration of the importance of realistic individualized treatment rules in causal inference. *Electronic Journal of Statistics*, 1:574–596, 2007.

O. Bembom and M.J. van der Laan. Data-adaptive selection of the truncation level for inverse-probability-of-treatment-weighted estimators. Technical Report 230, Division of Biostatstics, University of California, Berkeley, 2008. URL `www.bepress.com/ucbbiostat/paper230/`.

O. Bembom, J.W. Fessel, R.W. Shafer, and M.J. van der Laan. Data-adaptive selection of the adjustment set in variable importance estimation. Technical Report 231, Division of Biostatstics, University of California, Berkeley, 2008. URL `www.bepress.com/ucbbiostat/paper231/`.

O. Bembom, M.L. Petersen, S.Y. Rhee, W. J. Fessel, S.E. Sinisi, R.W. Shafer, and M.J. van der Laan. Biomarker discovery using targeted maximum likelihood estimation: Application to the treatment of antiretroviral resistant HIV infection. *Statistics in Medicine*, 28:152–72, 2009.

J. Bhattacharya and W. Vogt. Do instrumental variables belong in propensity scores? NBER Technical Working Paper 343, National Bureau of Economic Research, MA., 2007.

L. Breiman. Bagging predictors. *Machine Learning*, 24:123–140, 1996.

L. Breiman, J. H. Friedman, R. Olshen, and C. J. Stone. *Classification and regression trees*. The Wadsworth statistics/probability series. Wadsworth International Group, 1984.

W. Cao, A.A. Tsiatis, and M. Davidian. Improving efficiency and robustness of the doubly robust estimator for a population mean with incomplete data. *Biometrika*, 96,3:723–734, 2009.

C.C. Chang and C.J. Lin. Libsvm: a library for support vector machines (version 2.31). Technical report, 2001.
`http://www.csie.ntu.edu.tw/ cjlin/papers/libsvm2.ps.gz`.

W.G. Cochran. Analysis of covariance: Its nature and uses. *Biometrics*, 13:261–281, 1957.

S.R. Cole and M.A. Hernan. Constructing inverse probability weights for marginal structural models. *American Journal of Epidemiology*, 168:656–664, 2008.

C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20:273–297, December 1995.

R.K. Crump, V.J. Hotz, G.W. Imbens, and O.A. Mitnik. Moving the goalposts: Adressing limited overlap in the estimation of average treatment effects by changing the estimand. Technical Report 330, National Bureau of Economic Research, 2006. URL `http://www.nber.org/papers/T0330`.

R. Dehejia and S.Wahba. Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of the American Statistical Association*, 94:1053–1062, 1999.

E. Dimitriadou, K. Hornik, F. Leisch, D. Meyer, and A. Weingessel. *e1071: Misc Functions of the Department of Statistics (e1071), TU Wien*, 2010. URL `http://CRAN.R-project.org/package=e1071`. R package version 1.5-24.

B. Efron, Hastie H., Johnstone, and Tibshirani. Least angle regression (with discussion). *Annals of Statistics*, 2003.

D.A. Freedman and R.A. Berk. Weighting regressions by propensity scores. *Evaluation Review*, 32(4):392–409, 2008.

J.H. Friedman. Multivariate adaptive regression splines. *The Annals of Statistics*, 19(1):pp. 1–67, 1991. ISSN 00905364. URL `http://www.jstor.org/stable/2241837`.

J.H. Friedman. Fast mars. Technical report, Department of Statistics, Stanford University, 1993.

J.H. Friedman. Flexible metric nearest neighbor clasification. Technical report, Department of Statistics, Stanford University, 1994.

S. Greenland. Model-based estimation of relative risks and other epidemiologic measures in studies of common outcomes and in case-control studies. *Am J Epidemiol*, 160(4):301–305, 2003.

S. Gruber and M.J. van der Laan. A targeted maximum likelihood estimator of a causal effect on a bounded continuous outcome. Technical Report 265, UC Berkeley, 2010a.

S. Gruber and M.J. van der Laan. An application of collaborative targeted maximum likelihood estimation in causal inference and genomics. *The International Journal of Biostatistics*, 6,1(18), 2010b.

T. Hastie, R. Tibshirani, and J Friedman. *The Elements of Statistical Learning*. Springer, London, 2009.

T.J. Hastie and D. Pregibon. Generalized linear models. In J. M. Chambers and T. J. Hastie, editors, *Statistical Models in S*, chapter 6. Wadsworth & Brooks/Cole, 1992.

J. Heckman, H. Ichimura, and R. Todd. Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme. *Review of Economic Studies*, 64: 605–654, 1997.

C.J. Hogue, D.W Gaylor, and K.F. Schulz. Estimators of relative risk for case-control studies. *Am J Epidemiol*, 118(3):396–407, Sep 1983.

A. Rotnitzky J.M. Robins and L.P. Zhao. Estimation of regression coefficients when some regressors are not always observed. *J. Am. Statist. Assoc.*, 89:846–66, 1994.

V.A. Johnson, F. Brun-Vezinet, and et. al. B. Clotet. Update of the drug resistance mutations in HIV-1: December 2009. *Topics in HIV Medicine*, 17(5):138–45, 2009.

J. Kang and J. Schafer. Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data (with discussion). *Statistical Science*, 22:523–39, 2007.

L. Kish. Weighting for unequal p_i. *Journal of Official Statistics*, 8:183–200, 1992.

X. Xue L. McNutt, H. Chuntao and P. Hafner. Estimating the relative risk in cohort studies and clinical trials of common outcomes. *Am J Epidemiol*, 157(10):940–943, 2003.

R.J. LaLonde. Evaluating the econometric evaluations of training programs with experimental data. *American Economic Review*, 76:604–620, 1986.

Kronmal R. Lumley, T. and S. Ma. Relative risk regresssion in medical research: Models, contrasts, estimators, and algorithms. Technical Report 293, University of Washington, 2006.

S Milborrow. *earth: Multivariate Adaptive Regression Spline Models*, 2009. URL `http://CRAN.R-project.org/package=earth`. R package version 2.4-0.

K.L. Moore and M.J. van der Laan. Covariate adjustment in randomized trials with binary outcomes. Technical report 215, Division of Biostatistics, University of California, Berkeley, April 2007.

K.L. Moore, R.S. Neugebauer, M.J. van der Laan, and I.B. Tager. Causal inference in epidemiological studies with strong confounding. Technical Report 255, Division of Biostatistics, University of California, Berkeley, 2009. URL `www.bepress.com/ucbbiostat/paper255/`.

R. Neugebauer and M. J. van der Laan. Non-parametric causal effects based on marginal structural models. *Journal of Statistical Planning and Inference*, 137(2):419–434, 2007.

R. Neugebauer and M.J. van der Laan. Why prefer double robust estimates. *Journal of Statistical Planning and Inference*, 129(1-2):405–26, 2005.

Romain Neugebauer and James Bullard. *DSA: Deletion/Substitution/Addition algorithm*, 2010. URL `http://www.stat.berkeley.edu/ laan/Software/`. R package version 3.1.4.

J. Neyman. On the application of probability theory to agricultural experiments. essay on principles. section 9. *Statistical Science*, 5:465–480, 1923.

J. Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, Cambridge, 2000.

J. Pearl. On a class of bias-amplifying variables that endanger effect estimates. *Proceedings of UAI*, 2010. forthcoming.

A. Peters and T. Hothorn. *ipred: Improved Predictors*, 2009. URL `http://CRAN.R-project.org/package=ipred`. R package version 0.8-8.

M.L. Petersen, K. Porter, S. Gruber, Y. Wang, and M. van der Laan. Diagnosing and responding to violations in the positivity assumption. Technical report, Division of Biostatistics, Universtiy of California, Berkeley, 2010. URL `www.bepress.com/ucbbiostat/paperxxx/`.

G. Ridgeway and D. McCaffrey. Comment: Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data (with discussion). *Statistical Science*, 22:540–43, 2007.

B. D. Ripley. *Pattern recognition and neural networks*. Cambridge University Press, Cambridge, New York, 1996.

J.M. Robins. A new approach to causal inference in mortality studies with sustained exposure periods - application to control of the healthy worker survivor effect. *Mathematical Modelling*, 7:1393–1512, 1986.

J.M. Robins. Addendum to: "A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect" [Math. Modelling **7** (1986), no. 9-12, 1393–1512; MR 87m:92078]. *Comput. Math. Appl.*, 14(9-12):923–945, 1987a. ISSN 0097-4943.

J.M. Robins. A graphical approach to the identification and estimation of causal parameters in mortality studies with sustained exposure periods. *Journal of Chronic Disease*, 40(2): 139s–161s, 1987b.

J.M. Robins. Marginal structural models. In *Proceedings of the American Statistical Association. Section on Bayesian Statistical Science 1997*, pages 1–10, 1998.

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

J.M. Robins. Commentary on using inverse weighting and predictive inference to estimate the effecs of time-varying treatments on the discrete-time hazard. *Statistics in Medicine*, (21):1663–1680, 1999.

J.M. Robins. Marginal structural models versus structural nested models as tools for causal inference. In *Statistical models in epidemiology, the environment, and clinical trials (Minneapolis, MN, 1997)*, pages 95–133. Springer, New York, 1999.

J.M. Robins. Commentary on using inverse weighting and predictive inference to estimate the effects of time-varying treatments on the discrete-time hazard by Dawson and Lavori". *Statistics in Medicine*, 21:1663–1680, 2002.

J.M. Robins and A. Rotnitzky. Comment on the Bickel and Kwon article, "Inference for semiparametric models: Some questions and an answer". *Statistica Sinica*, 11(4):920–936, 2001.

J.M. Robins and N. Wang. Inference for imputation estimators. *Biometrika*, 87:113–124, 2000.

J.M. Robins, A. Rotnitzky, and L.P. Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89 (427):846–66, September 1994.

J.M. Robins, M.A. Hernan, and B. Brumback. Marginal structural models and causal inference in epidemiology. *Epidemiology*, 11(5):550–560, 2000.

J.M. Robins, M. Sued, Q. Lei-Gomez, and A. Rotnitzky. Comment: Performance of double-robust estimators when "inverse probability" weights are highly variable. *Statistical Science*, 22:544–559, 2007.

J.M. Robins, L. Orellana, and Andrea Rotnitzky. Estimation and extrapolation of optimal treatment and testing strategies. *Statistics in Medicine*, 27:4678–4721, 2008.

S. Rose and M.J. van der Laan. Simple optimal weighting of cases and controls in case-control studies. *The International Journal of Biostatistics*, page http://www.bepress.com/ijb/vol4/iss1/19, 2008.

S. Rose and M.J. van der Laan (Eds.). *Targeted Learning: Causal Effect Estimation in Observational and Experimental Studies*. Springer, New York, NY, 2011.

M. Rosenblum and M. van der Laan. Targeted maximum likelihood estimation of the parameter of a marginal structural model. *The International Journal of Biostatistics*, 6(2), 2010a.

M. Rosenblum and M.J. van der Laan. Confidence intervals for the population mean tailored to small sample sizes, with applications to survey sampling. *The International Journal of Biostatistics*, 1:4, 2001.

M. Rosenblum and M.J. van der Laan. Targeted maximum likelihood estimation of the parameter of a marginal structural model. *The International Journal of Biostatistics*, 6 (19), 2010b.

M. Rosenblum and M.J. van der Laan. Simple, efficient estimators of treatment effects in randomized trials using generalized linear models to leverage baseline variables. *The International Journal of Biostatistics*, 6(1,13), 2010.

D.B. Rubin. Estimating causal effects of treatments in randomized and non-randomized studies. *Journal of Educational Psychology*, 66:688–701, 1974.

D.B. Rubin and M.J. van der Laan. Empirical efficiency maximization: Improved locally efficient covariate adjustment in randomized experiments and survival analysis. *The International Journal of Biostatistics*, Vol. 4, Iss. 1, Article 5, 2008.

D.O. Scharfstein, A. Rotnitzky, and J.M. Robins. Adjusting for non-ignorable drop-out using semiparametric nonresponse models, (with discussion and rejoinder). *Journal of the American Statistical Association*, (94):1096–1120 (1121–1146), 1999.

S.E. Sinisi and M.J. van der Laan. Deletion/Substitution/Addition algorithm in learning with applications in genomics. *Statistical Applications in Genetics and Molecular Biology*, 3(1), 2004. URL `http://www.bepress.com/sagmb/vol3/iss1/art18`. Article 18.

Z. Tan. A distributional approach for causal inference using propensity scores. *J. Am. Statist. Assoc.*, 101:1619–37, 2006.

Z. Tan. Comment: Understanding or, ps and dr. *Statistical Science*, 22:4:560–568, 2007.

Z. Tan. Bounded, efficient and doubly robust estimation with inverse weighting. *Biometrika*, 97,3:661–682, 2010.

R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2010.

A. Tsiatis and M. Davidian. Comment: Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data (with discussion). *Statistical Science*, 22:569–73, 2007.

C. Tuglus, K.E. Porter, and M.J. van der Laan. Targeted maximum likelihood estimation of conditional relative risk parameters in a semi-parametric multiplicative regression model. *in preparation*, 2011.

M.J. van der Laan. Estimation based on case-control designs with known prevalance probability. *The International Journal of Biostatistics*, page http://www.bepress.com/ijb/vol4/iss1/17/, 2008.

M.J. van der Laan and S. Dudoit. Unified cross-validation methodology for selection among estimators and a general cross-validated adaptive epsilon-net estimator: Finite sample oracle inequalities and examples. Technical report, Division of Biostatistics, University of California, Berkeley, November 2003a.

M.J. van der Laan and S. Dudoit. Unified cross-validation methodology for selection among estimators and a general cross-validated adaptive epsilon-net estimator: Finite sample oracle inequalities and examples. Technical Report 130, Division of Biostatistics, University of California, Berkeley, 2003b. URL `www.bepress.com/ucbbiostat/paper130/`.

M.J. van der Laan and S. Gruber. Collaborative double robust targeted penalized maximum likelihood estimation. Technical Report 246, Division of Biostatistics, University of California, Berkeley, 2009a. URL `www.bepress.com/ucbbiostat/paper246/`.

M.J. van der Laan and S. Gruber. Collaborative double robust penalized targeted maximum likelihood estimation. *The International Journal of Biostatistics*, 2009b.

M.J. van der Laan and M.L. Petersen. Causal effect models for realistic individualized treatment and intention to treat rules. *International Journal of Biostatistics*, 3(1), 2007.

M.J. van der Laan and J.M. Robins. *Unified methods for censored longitudinal data and causality*. Springer, New York, 2003.

M.J. van der Laan and D. Rubin. Targeted maximum likelihood learning. *The International Journal of Biostatistics*, 2(1), 2006.

M.J. van der Laan, S. Dudoit, and A.W. van der Vaart. The cross-validated adaptive epsilon-net estimator. Technical report 142, Division of Biostatistics, University of California, Berkeley, February 2004a.

M.J. van der Laan, A.E. Hubbard, and N.P. Jewell. Estimation of treatment effects in randomized trials with noncompliance and a dichotomous outcome. Technical report 157, Division of Biostatistics, University of California, Berkeley, September 2004b.

M.J. van der Laan, E.C. Polley, and A. Hubbard. Super learner. *Statistical Applications in Genetics and Molecular Biology*, 6(25), 2007. ISSN 1.

M.J. van der Laan, S. Rose, and S. Gruber. Readings on targeted maximum likelihood estimation. *Technical report, working paper series http://www.bepress.com/ucbbiostat/paper254*, 2009.

A. W. van der Vaart and J. A. Wellner. *Weak Convergence and Emprical Processes*. Springer-Verlag New York, 1996.

W.N. Venables and B.D. Ripley. *Modern applied statistics with S*. Springer, New York, 4th edition, 2002.

Y. Wang, M. Petersen, D. Bangsberg, and M.J. van der Laan. Diagnosing bias in the inverse probability of treatment weighted estimator resulting from violation of experimental treatment assignment. Technical Report 211, Division of Biostatistics, University of California, Berkeley, 2006a.

Y. Wang, M. Petersen, and M.J. van der Laan. A statistical method for diagnosing eta bias in iptw estimators. Technical report, Division of Biostatistics, University of California, Berkeley, 2006b.

R.W.M. Wedderburn. Quasi-likelihood functions, generalized linear models, and the gauss-newton method. *Biometrika*, 61, 1974.

J. Wooldridge. Should instrumental variables be used as matching variables? Tech. Rep. Michigan State University, MI., 2009.

E. Grosse W.S. Cleveland and W.M. Shyu. Local regression models. In J. M. Chambers and T. J. Hastie, editors, *Statistical Models in S*, chapter 6. Wadsworth & Brooks/Cole, 1992.