Title
Targeted Minimum Loss Based Estimation for Longitudinal Data

Author
Chaffee, Paul H.

Publication Date
2012

Peer reviewed|Thesis/dissertation

Targeted Minimum Loss Based Estimation for Longitudinal Data

by

Paul H. Chaffee

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 van der Laan, Chair
Professor Alan Hubbard
Professor Jasjeet Sekhon

Fall 2012

Abstract

Targeted Minimum Loss Based Estimation for Longitudinal Data

by

Paul H. Chaffee

Doctor of Philosophy in Biostatistics

University of California, Berkeley

Professor Mark J. van der Laan, Chair

Sequential Randomized Controlled Trials (SRCTs) are rapidly becoming essential tools in the search for optimized treatment regimes in ongoing treatment settings. Analyzing data for multiple time-point treatments with a view toward optimal treatment regimes is of interest in many types of afflictions: HIV infection, Attention Deficit Hyperactivity Disorder in children, leukemia, prostate cancer, renal failure, and many others. Methods for analyzing data from SRCTs exist but they are either inefficient or suffer from the drawbacks of estimating equation methodology. This dissertation describes the development of a general methodology for estimating parameters that would typically be of interest both in SRCTs and in observational studies which are longitudinal in nature, and have multiple time-point exposures or treatments. It is expected in such contexts that time-dependant confounding is either present (observational studies) or actually designed in as part of a study (SRCTs). The method, targeted minimum loss based estimation (TMLE), has been fully developed and implemented in point treatment settings and for various outcome types, including time to event outcomes, and binary and continuous outcomes. Here we develop and implement TMLE in the longitudinal setting, and pay special attention to dynamic treatments or exposures, as might be seen in SRCTs. Dynamic exposures are not limited to SRCTs however. The idea of a rule-based intervention turns out be a very fruitful one when one faces complex treatment or exposure patterns, or when one encounters challenges in defining an intervention that must depend on time-varying factors. As in the former settings, the TMLE procedure is targeted toward a pre-specified parameter of the distribution of the observed data, and thereby achieves important bias reduction over non-targeted procedures in estimation of that parameter. As with the so-called Augmented Inverse Probability of Censoring Weight (A-IPCW) estimator, TMLE is double-robust and locally efficient. We develop some of the background involving the causal and statistical models and report the results of several simulation studies under various data-generating distributions and for two outcome types (binary, and continuous on [0,1]). In our results we include comparisons from a number of other estimators in current use.

Chapter 1 develops the background and context in which this estimator appears, gives a brief history of other estimators used in SRCTs and describes some of the theory behind TMLE in the longitudinal setting. Two different TMLE algorithms are described in detail, and results of a simulation study for

three separate causal parameters are presented.

Chapter 2 concerns the development of a new TMLE that solves the efficient influence curve estimating equation directly by numerical methods, rather than indirectly, which is the usual procedure. A new set of simulations is performed here that compare this TMLE with the preceding two (presented in chapter 1). Its performance is comparable to those described in chapter 1, but it is somewhat easier to implement.

Chapter 3 is a comparison of still another new TMLE (described in van der Laan and Gruber, 2012) with one of the three described above. This TMLE arguably shows the most promise generally, since it's implementation does not require discretization of the intermediate factors of the likelihood as does the three preceding TMLEs. Further, under the right conditions it exhibits superior performance in terms of MSE. We also explore a new, targeted criterion for selecting the initial estimators involved.

Chapter 4 describes a detailed analysis of the estimation of the effect of gestational weight gain on women's long term BMI using the preferred TMLE described in chapter 3. Many issues were encountered during this analysis concerning censoring of the exposure variable that led to the redefinition of the parameter of interest, and the implementation of a different type of TMLE for the first time (described originally in van der Laan, 2008). We also encountered issues arising from sparsity in the data and propose and implement corresponding solutions. The analysis was performed using data from the national longitudinal survey of youth, begun in 1979 and ending in 2008.

To Elina and Laszlo

Contents

Acknowledgements

I have always had great luck in encountering, unplanned, brilliant thinkers to study under in my academic career. In the graduate program in biostatistics at Berkeley, that previous lucky streak increased beyond all plausibility.

I am indebted first and foremost to my advisor, Professor Mark van der Laan. Without the tremendous work he's done paving the way in the realm of efficient estimation of the type of parameters that are so prevalent in the area of public health, the field would be a much poorer one. The novelty, depth and scope of his ideas have not only enriched the field, but have opened up numerous new paths of research that his students, myself included, can now take. It is a genuine honor to have worked under his direction and guidance, and to have experienced his genius at such close range.

I would also like to thank Professors Alan Hubbard and Nick Jewell. Their guidance over the years has been superb, their ideas and approach to the field of biostatistics, novel and inspirational. I am also thankful to Maya Peterson for contributing much to solidifying some of the foundations of my knowledge in the area of causal inference. They have very different styles of doing statistics, and it's been a pleasure to try to incorporate each into my own approach.

Sharon Norris has been an advocate for me in many ways during the years and I'm indebted to her for all she's done to help me navigate the university bureaucracy. She has generously placed me in line for beneficial opportunities I would not otherwise have been considered for.

Other students in the doctoral program have also been a great help and inspiration to me, and have contributed to making life during the program more of a pleasure than a burden. Among them are Iván Díaz, Molly Davies, Inna Gerlovina, Stephan Ritter, Oliver Bembom, Wenjing Zheng, Curt Hansen, Farid Jamshidian, Sam Lendle, Luca Pozzi, Sherri Rose, Ori Stitelman, Susan Gruber, Brian Greenhouse, and at least a few others who've contributed to my development along the way.

My wife, Elina Coulter, has exhibited superhuman patience in allowing the process to unfold, not least because pressing ahead for a Ph.D after completing the Masters was never part of the original plan. Her support along the way has sustained me, and to say that she helped me move past the low points is an understatement. I don't know that I could have done it without her.

My beloved son Laszlo was born in the middle of my first year of the Master's program, and he too never gave up on me (at least as far as I can tell). It was always a welcome and necessary diversion to spend time with him when I needed a break from my work.

Finally, it was worth it alone to complete this dissertation just to get Ken Bouley and Elliott Long off my back about it.

Chapter 1

Targeted Minimum Loss Based Estimation with Application to Sequentially Randomized Controlled Trials with Dynamic Treatment Rules

1.1 Introduction

The treatment of many types of afflictions involves ongoing therapy—that is, application of therapy at more than one point in time. Therapy in this context often involves treatment of patients with drugs, but need not be limited to drugs. For example, the use of pill organization devices ("pillboxes") has been studied as a means to improve drug adherence (Petersen et al., 2007), and others (Moodie et al., 2009) have studied the optimum time at which infants should stop breastfeeding.

A common setting for ongoing treatment therapy involves randomization to initial treatment (or randomization to initial treatment within subgroups of the population of interest), followed by later treatments which may also be randomized, or randomized to a certain subset of possible treatments given that certain intermediate outcomes occurred, by definition, after the initial treatment. Examples from the literature include treatment by antipsychotic medications for reduction in severity of schizophrenia symptoms (Tunis et al., 2006), treatment of prostate cancer by a sequence of drugs determined by success or failure of first-line treatment (Bembom and van der Laan, 2007), when HIV patients should switch treatments (Orellana et al. 2010, van der Laan and Petersen 2007) and many others.

Suppose, for example, that every subject in a prostate cancer study is randomized to an initial pair of treatments (A or B, say), and if a subject's tumor size increases or does not decrease, the subject is again randomized to A or B at the second treatment point. On the other hand, if the subject does well on the first treatment (tumor size decreases, say), then he or she is assigned the same treatment at the second time point as the first. The general term for multiple time point treatments in which treatments after the first-line are assigned in response to intermediate outcomes is *dynamic treatment regimes* or *dynamic treatment rules* (Murphy et al., 2001). If the intermediate outcome in such SRCTs is affected by initial treatment, and in turn affects decisions at the second time-point treatment as well as the final outcome, then it is a so-called "time-dependent confounder."

1.1.1 Existing Procedures

A number of methods have been proposed to estimate parameters associated with such a study. This article describes implementation of targeted maximum likelihood estimation for two time-point longitudinal data structures, and is based on the framework developed for general longitudinal data structures presented in van der Laan (2010a,b).

Lunceford et al. (2002) develop inverse probability of treatment weighted (IPTW) estimators and an estimating equation estimator suitable for analysis of survival times from a leukemia clinical trial. Wahed and Tsiatis (2004) propose an estimating equation-based estimator which uses the efficient influence curve for estimating treatment policy-specific parameters in two-stage clinical trials. They later extended those methods to account for right-censoring in such trials (Wahed and Tsiatis, 2006). Guo and Tsiatis (2005) develop what they call a "Weighted Risk Set Estimator" for use in two-stage trials where the outcome is a time-to-event (such as death). Tunis et al. (2006) use IPTW methods, Marginal Structural Models and the so-called "g-estimation" method for analyzing the causal effect of a "continuous" treatment regime of atypical antipsychotic medications on severity of schizophrenia symptoms. This study/analysis involved no time-dependent confounders, however. Laber et al. (2009) use Q-learning to estimate optimal dynamic treatment regimes in Attention Deficit Hyperactivity Disorder in children. Miyahara and Wahed (2010) used weighted Kaplan-Meier estimators for estimating treatment-specific survival rates. Orellana et al. (2010) use structural marginal mean models, IPTW and the so-called augmented inverse probability of censoring weighted (A-IPCW) estimators with a view toward estimating optimal treatment regimes for switching to HAART therapy among HIV-positive patients. Bembom and van der Laan (2007) apply simple g-computation and IPTW estimation procedures in analyzing the optimum response of prostate cancer patients to randomized first-line treatment followed by second-line treatment

which was either 1) the same as the first line treatment if that had been deemed successful, or 2) randomized to three remaining treatments if the first line had failed. The data used for the latter analysis has recently been re-analyzed using stabilized IPTW estimation by Wang et al. (2012). The latter article was the subject of discussion articles, among them a general presentation of the methods described here (Chaffee and van der Laan, 2012). This type of trial and data closely resembles what we simulate and analyze in the present study, though we add baseline covariates and more than two levels of success in the intermediate biomarker covariate in order to generalize the data structure to more types of scenarios.

We present a new estimator for this longitudinal data structure: the targeted maximum likelihood estimator (TMLE). TMLE has application in a wide range of data structures and sampling designs (van der Laan and Rose, 2011). Though this estimator can be applied to a broad range of data structures of longitudinal type, we focus here on the estimation of treatment-rule-specific mean outcomes. This also covers static treatment regimes for the given data structures.

In the next section we describe the data structure and define the likelihood for the scenarios we intend to analyze. Once we have specified a counterfactual target parameter of interest and equated it with a well-defined mapping from conditional distributions of the data to a real number, we describe TMLE in broad outline, and in particular, the implementation of two different estimators grounded in the general TMLE approach. Specifically we present the so-called efficient influence curve for certain parameters of interest and show the relationship between elements of this object and elements of the targeted maximum likelihood estimators. Following these general descriptions we present simulation results, including details of specific treatment rules, data generation and results in terms of bias, variance and relative mean squared error. A short discussion of the results follows.

1.2 Data Structure and Likelihood

In the settings of interest here, a randomly sampled subject has data structure $O = (L(0), A(0), L(1), A(1), Y = L(2)) \sim P_0$, where $L(0)$ indicates a vector of baseline covariates, $A(0)$ is initial randomized treatment, $L(1)$ is, say, an intermediate biomarker (which we first consider as binary), $A(1)$ is the second time point treatment (which we also take as binary), $Y = L(2)$ is the clinical outcome of interest and P_0 is the joint distribution of O. We take the data to be n *i.i.d.* copies of O. We also assume $A(1)$ can be set in response to $L(1)$. The patient's full treatment is therefore $(A(0), A(1))$, and specific realizations of $(A(0), A(1))$ may or may not constitute realizations of a specific dynamic treatment rule. Such "rules" are dynamic in the sense that the regimen can

be set according to a patient's response to treatment over time. However, even if $A(0)$ and $A(1)$ are both unconditionally randomized, parameters of the distribution of the above data can nevertheless be identified which correspond with dynamic treatment regimens.

The data structure for such an experimental unit can be thought of as a time series in discrete time. For many of the (not necessarily regularly-spaced) time points there may be no observation of interest, and at others measurable events of interest occur. Many measurable events may occur at the same time—e.g., assignment of treatment and recording of measured characteristics. A specified set of all measured variables that respects this time-ordering, together with possible additional knowledge about the ordering and relationships of the variables, implies a particular statistical graph. The graph is a representation of each variable and its causal relation to its parent nodes, the latter being defined as all variables that preceded it in the specified time-ordering and are either direct or indirect causal antecedents. The graph can be modified to encode not only the time-ordering of the variables but also possible additional causal assumptions. The likelihood of this unit-specific data structure can be factorized according to the specified time-ordering, where the factors consist of the conditional distribution of each node given its parents, for all nodes in the graph.

The likelihood of the data described above can be factorized as

$$p(O) = \prod_{j=0}^{2} P\big(L(j) \mid \bar{L}(j-1), \bar{A}(j-1)\big) \prod_{j=0}^{1} P\big(A(j) \mid \bar{L}(j), \bar{A}(j-1)\big), \quad (1.1)$$

where $\bar{A}(j) = (A(0), A(1), ..., A(j))$ and $\bar{L}(j)$ is similarly defined. Factorizing the likelihood in this way is suggested by the time–ordering of the variables in O. That is, we assume $L(0)$ is followed by $A(0)$, and then $L(1)$, $A(1)$ and outcome $L(2)$ occur in that order. The above formula is the most general in the sense that each factor is represented as a function of its parents as defined by the time-ordering of the data, but in some cases a particular factor may be a function of fewer nodes than this representation suggests. (An example is given later in this section.)

Equation (4.1.1) is an example of the general longitudinal factorization

$$p_0(O) = \prod_{k=1}^{K} P\left(N(k) \mid Pa(N(k))\right),$$

where $N(k)$ denotes node k, corresponding to observed variable k in the graph, and $Pa(N(k))$ are the parents of $N(k)$ (van der Laan, 2010a). We make no

assumptions on the conditional distributions of $N(k)$ for each $k = 0, 1, 2...K$ beyond $N(k)$'s depending only on $Pa(N(k))$.

For simplicity, we introduce the notation $Q_{L(j)}$, $j = 0, 1, 2$ to denote the factors of (4.1.1) under the first product and $g_{A(j)}$, $j = 0, 1$ for those under the second; the latter we refer to as the *treatment and/or censoring mechanism.* Thus in the simpler notation we have

$$p(O) = \prod_{j=0}^{2} Q_{L(j)} \prod_{j=0}^{1} g_{A(j)} = Qg.$$

The factorization of the likelihood alone puts no restrictions on the possible set of data-generating distributions, but does affect the so-called G-computation formula for the counterfactual distributions of the data under any interventions implied by the ordering. The G-computation formula also specifies the set of nodes on which to intervene, as well as the interventions that correspond to the parameter of interest. For the data structures of interest here, interventions will be on the treatment nodes $(A(0), A(1))$. These interventions could be simply static assignment of treatment at each time point, or the above-mentioned dynamic treatment rules.

A typical parameter of interest in point treatment settings is the treatment-specific mean. For example if A is treatment, with levels $a = \{0, 1\}$, a causal parameter of interest might be EY_1, which is the mean outcome of the population had that entire population received treatment 1. Similarly, we define a treatment-specific mean for the multiple time point data structure where now a particular treatment means a specific treatment course over time. We define a *treatment rule*, $d = (d_0, d_1)$ for the treatment points $(A(0), A(1))$, which is the set of mappings $d_0 : \mathcal{D}_0 \to \mathcal{A}_0, d_1 : \mathcal{D}_1 \to \mathcal{A}_1$, where $\mathcal{A}_j, j = 0, 1$ is the set of possible values for $A(j)$, $\mathcal{D}_0$ is the support of $L(0)$ and $\mathcal{D}_1$ is the support of $(L(0), A(0), L(1))$. We can express the overall rule as $d(\bar{L}(1)) = \big(d_0(L(0)), d_1(\bar{L}(1))\big)$. Under this definition we can easily express either static or dynamic treatment rules, or a combination of the two (see examples in section 1.4.1). For example, $d_0 = 1$ would correspond to a static assignment for $A(0)$, and $d_1 = I(L(1) = 1) * 1 + I(L(1) = 0) * 0$ is dynamic since it assigns treatment $A(1)$ in response to the patient's intermediate outcome, $L(1)$.

We can now define the G-formula to be the product across all nodes, excluding intervention nodes, of the conditional distribution of each node given its parent nodes, and with the values of the intervention nodes fixed according to the static or dynamic intervention of interest. This formula thus expresses the distribution of $\bar{L}$ given $\bar{A} = (A(0), A(1))$ is at value $d(\bar{L})$.

$$P^d(\bar{L}) = \prod_{j=0}^{2} Q^d_{L(j)}(\bar{L}(j)), \tag{1.2}$$

where we used the notation

$$Q^d_{L(j)}(\bar{L}(j)) \equiv P(\bar{L}(j) \mid \bar{L}(j-1), \bar{A}(j-1) = d(\bar{L}(j-1))).$$

The superscript d here denotes that the joint distribution of $\bar{L}$ is conditional on $\bar{A} = d(\bar{L})$. We reserve subscript d to refer to counterfactually-defined variables.

Under the right conditions on the causal graph augmented by a set of nodes that include unobserved variables (see below), the G-computation formula equals the counterfactual distribution of the data had one carried out the specified intervention described by the graph. In point treatment settings the conditions are desribed as *no unblocked backdoor paths from intervention node to outcome node*, or in alternative formulation, *d-separation of intervention and outcome nodes conditional on some subset of observed nodes* (Pearl, 2000). Meeting these assumptions typically implies meeting the so-called randomization assumption. In longitudinal settings, the analog is the sequential randomization assumption (SRA) which is a generalized version of the *no unblocked backdoor path* condition, applied to multiple treatment nodes, defined formally below.

1.2.1 Causal and Statistical Models

We signify the non-parametric causal model of interest $\mathcal{M}^{\mathcal{F}}$, which includes all possible distributions compatible with a specified causal structure. Such a structure can be encoded in the form of an acyclic graph as mentioned above, or a set of structural equations. The set of such equations, together with possible additional causal assumptions defines a so-called structural causal model (SCM). Restrictions on relationships between nodes (other than those implied by the time ordering itself) can reduce the size of the set of parent nodes for a given node, and result in a semi-parametric causal model. The non-parametric set of such equations (i.e., with no exclusion restrictions) corresponding to the data structure here, for example, is

$U = (U_{L(0)}, U_{A(0)}, U_{L(1)}, U_{A(1)}, U_Y) \sim P_U$
$L(0) = f_{L(0)}\left(U_{L(0)}\right)$
$A(0) = f_{A(0)}\left(L(0), U_{A(0)}\right)$
$L(1) = f_{L(1)}\left(L(0), A(0), U_{L(1)}\right)$

$$A(1) = f_{A(1)}\left(L(0), A(0), L(1), U_{A(1)}\right)$$
$$Y = f_Y\left(L(0), A(0), L(1), A(1), U_Y\right),$$

where $U_{L(0)}, U_{A(0)}$, etc., are the so-called exogenous variables of the system—random inputs associated with each of the graph nodes that are not affected by any other variable in the model. The SCM represented above does not restrict the set of functions $F = \left\{f_{L(0)}, f_{A(0)}, ...f_Y\right\}$ to any particular functional form. Further, each node is represented as a function of the complete set of parent nodes implied by the time ordering. If, in addition, no assumptions are made about the independence of the variables in U, then the causal model is fully non-parametric. (This formulation of the SCM is based on Pearl, 2000.)

The nodes in the graph correspond to the endogenous variables—those variables that are affected by other variables in the graph, which we denote generically as $X = \{X_1, ...X_J\}$. For the SCM depicted above, the set X consists of the observed variables, i.e., $X = O$. Each endogenous variable, X_j, is the solution of a deterministic function of its parents and U_j; the latter represents all the unknown mechanisms that are involved in the generation of X_j. The causal model can now be expressed as all probability distributions compatible with the SCM. Elements of the *observed* data model, $\mathcal{M}$, can be thought of as being indexed by the elements of $\mathcal{M}^{\mathcal{F}}$, i.e., for every P in $\mathcal{M}$, $P = P_{P_{U,X}}$ for some $P_{U,X} \in \mathcal{M}^{\mathcal{F}}$, or, alternatively, $\mathcal{M} = \left\{P_{P_{U,X}} : P_{U,X} \in \mathcal{M}^{\mathcal{F}}\right\}$.

Assumptions of independence between any of the $U's$ have implications for identifiability of the causal parameter in terms of the distribution of the observed data. For example, strict randomization of $A(0)$ makes $U_{A(0)}$ independent of all other $U's$, which will typically reduce the number of additional assumptions needed for identifiability. Excluding nodes from the parent set of a given node restricts the set of allowed distributions of the observed data, $\mathcal{M}$, corresponding to $\mathcal{M}^{\mathcal{F}}$.

Suppose now that we are interested in the outcomes of individuals had their treatment regimen been assigned according to some rule, d. Given a particular SCM such as the one defined above, we can write Y_d, the so-called counterfactual outcome under rule d, as the solution to the equation

$$Y_d = f_Y(L(0), A(0) = d_0(L(0)), L_d(1), A(1) = d_1(\bar{L}), U_Y),$$

where now $L_d(1)$ is the value $L(1)$ takes under rule d. The full SCM under intervention d is

$$U = (U_{L(0)}, U_{L(1)}, U_Y) \sim P_U$$
$$L(0) = f_{L(0)}(U_{L(0)})$$
$$A(0) = d_0(L(0))$$
$$L_d(1) = f_{L(1)}(L(0), A(0) = d_0(L(0)), U_{L(1)})$$

$A(1) = d_1(\bar{L})$
$Y_d = f_Y(L(0), d_0, L_d(1), d_1, U_Y).$

With the counterfactual outcome Y_d now defined in terms of the solution to a system of structural equations, we can define a corresponding counterfactual parameter of $P_{U,X}$, say $\Psi^F(P_{U,X}) = EY_d$, which in fact is the parameter we concern ourselves with in this article. Using (1.2),

$$\Psi^F(P_{U,X}) = EY_d = \sum_{l(0),l(1)} E\left(Y_d \mid L(0) = l(0), L_d(1) = l(1)\right) \prod_{j=0}^{1} Q_{L_d(j)}(\bar{l}(j)), \tag{1.3}$$

where $Q_{L_d(j)} \equiv P(L_d(j) \mid \bar{L}_d(j-1))$ and we omit the subscript d on $L(0)$ since it is prior to any treatment. In words, this parameter is the mean outcome under $P_{U,X}$ when treatment is set according to $\bar{A} = d(\bar{L})$.

As mentioned above, the parent set of nodes for any given node can be reduced if confirmed by additional knowledge of the conditional distribution of the node. If it is known, for example, that a particular node is a function only of a subset of its parents, then the parent nodes not in that subset can be excluded from the conditional distribution of that node. Such putative knowledge reduces the size of the model for the data-generating distribution, and can be tested from the data. For example, if $A(1)$ is assigned such that it is only a function of $L(1)$ then the set $Pa\left(A(1)\right) \backslash L(1)$ provides no information about the probability of $A(1)$ beyond that contained in $L(1)$, so

$$P\big(A(1) \mid Pa(A(1)\big) \equiv P\big(A(1) \mid L(0), A(0), L(1)\big) = P\big(A(1) \mid L(1)\big).$$

Once an SCM is committed to, one can formally state the assumptions on the SCM required in order for a particular G-computation formula for the *observed* nodes to be equivalent to the G-computation formula for the full set of nodes (1.3), which includes any relevant unobserved nodes. The latter can be viewed as the true causal parameter of interest (Pearl, 2000).

For the parameter of interest here, EY_d, the sequential randomization assumption (SRA), $Y_d \perp A(j) \mid Pa(A(j))$ for $j = 0, 1$, is sufficient for equivalence of the causal parameter $\Psi^F(P_{U,X})$ and a particular parameter of the observed data distribution $\Psi(P_0)$ for some Ψ (Robins, 1986). In particular, the SRA implies

$$\Psi^F(P_{U,X}) \equiv EY_d = \tag{1.4}$$
$$\Psi(P_0) = \sum_{l(0),l(1)} E\big(Y \mid L(0)=l(0), L(1)=l(1), \bar{A}=d(\bar{L})\big)\times$$
$$P\big(L(1)=l(1) \mid L(0)=l(0), A(0)=d_0\big)\times$$
$$P\big(L(0)=l(0)\big),$$

which is the so-called *identifiability result*.

Note that this parameter depends only on the Q part of the likelihood and we therefore also write $\Psi(P_0) = \Psi(Q_0)$. Note also that the first two factors in the summand are undefined if either $P\big(\bar{A}=d(\bar{L}) \mid L(0)=l(0), L(1)=l(1)\big)$ or $P\big(A(0)=d_0 \mid L(0)=l(0)\big)$ are 0 for any $(l(0), l(1))$, and so we require these two conditional probabilities to be positive. This is the so-called *positivity assumption*.

In this article we present a method for semi-parametric efficient estimation of causal effects. This is achieved through estimation of the parameters of the G-computation formula given above. The method is based on n independent and identically distributed observations of O, and our statistical model $\mathcal{M}$, corresponding to the causal model $\mathcal{M}^{\mathcal{F}}$, makes no assumptions about the conditional distribution of $N(k)$ given its parents, for each k in the graph.

Our parameter of interest, EY_d, can be approximated by generating a large number of observations from the intervened distribution P_d and taking the mean of the final outcome, in this case $L(2)$. The joint distribution P_d can itself be approximated by simulating sequentially from the conditional distributions $Q_{L_d(j)}, j=0,1,2$ to generate the observed values $L(j)$.

EY_d can also be computed analytically:

$$\begin{aligned}\Psi(Q_0) \equiv EY_d &= \sum_y y \sum_{l(0),l(1)} P_d\big(l(0),l(1),y\big)\\ &\overset{SRA}{=} \sum_y y \sum_{l(0),l(1)} P\big(Y=y \mid \bar{A}=d(\bar{L}), L(0)=l(0), L(1)=l(1)\big)\times\\ &\qquad P\big(L(1)=l(1) \mid L(0)=l(0), A(0)=d_0(L(0))\big] \times P[L(0)=l(0)\big)\\ &= \sum_y y \sum_{l(0),l(1)} Q^d_{L(2)}\big(l(0),l(1),y\big) Q^d_{L(1)}\big(l(0),l(1)\big) Q^d_{L(0)}\big(l(0)\big),\end{aligned}$$

The last expression is equivalent to the RHS of (1.4) if Y is binary. If $L(0)$ is continuous, the sum over $l(0)$ is replaced by an integral. The integral is replaced in turn by the empirical distribution if the expression above is approximated from a large number of observations. In that case the last line

reduces to

$$\Psi(Q_0) = \frac{1}{n}\sum_{i=1}^{n}\sum_{y}\sum_{l(1)} y\, Q^d_{L(2)}\big(L(0)_i, l(1), y\big) Q^d_{L(1)}\big(L(0)_i, l(1)\big). \tag{1.5}$$

The latter expression represents a well-defined mapping from the conditional distributions $Q_{L(j)}$ to the real line. Given an estimator $Q_n \equiv \prod_{j=0}^{2} Q_{L(j)_n}$ of $Q_0 \equiv \prod_{j=0}^{2} Q_{L(j)}$ we arrive at the substitution estimator $\Psi(Q_n)$ of $\Psi(Q_0)$.

Next we describe the targeted minimum loss based estimator (TMLE) of the relevant parameters of the G-computation formula. The TMLE is double-robust and locally efficient. The methods described here extend naturally to data structures with more time points, and/or more than one time-dependent confounder per time point (van der Laan, 2010a).

1.3 Targeted Maximum Likelihood Estimator

With the above parameter now established to be a well-defined mapping from the distribution of the data to the real line, we turn to the estimation of the conditional distributions, $Q_{L(j)}$ which are the domains of the function defining the parameter of interest, $\Psi(Q_0)$.

1.3.1 Basic Description

In targeted maximum likelihood estimation we begin by obtaining an initial estimator of Q_0; we then update this estimator with a fluctuation function that is tailored specifically to remove bias in estimating the particular parameter of interest. Naturally, this means that the fluctuation function is a function of the parameter of interest. There are, of course, various methods for obtaining an initial estimator: one can propose a parametric model for each factor $Q_{L(j)}$ and estimate the coefficients using maximum likelihood, or one can employ machine learning algorithms which use the data itself to build a model. The former method involves using standard software if the factors $L(j)$ are binary. Each of these general methods in turn has many variants. We favor machine learning, and in particular the Super Learner approach (van der Laan et al., 2007a). We recommend the latter approach in all cases because even if one feels one knows the true parametric model (and guessing the true model is highly unlikely) that belief can be validated by including this parametric model in the Super Learner library. If the model has good predictive results (where "good" here means low estimated cross-validated risk using an appropriate loss function) it will tend to be weighted highly in the final model returned by the Super Learner. If not, then the data do not support the analyst's

guess and the model will be given a low weight. Moreover, the authors of the Super Learner algorithm have shown that this particular machine learning approach yields a model whose asymptotic properties approach those of the "oracle" selector amongst the learners included in the Super Learner library. There thus appears to be nothing to lose—and everything to gain—in using this approach to obtaining an initial estimator $Q^{(0)}$ of Q_0. (Here we change notation slightly: the superscript (0) denotes the initial step in a multi-step algorithm, and does not signify a treatment rule.)

Upon obtaining an initial estimate $Q^{(0)}$ of Q_0, the next step in TMLE is to apply a fluctuation function to this initial estimator that is the least favorable parametric submodel through the initial estimate, $Q^{(0)}$ (van der Laan and Rubin, 2006). This parametric submodel through Q_0 is chosen so that estimation of $\Psi(Q_0)$ is "hardest in the sense that the parametric Cramer-Rao Lower Bound for the variance of an unbiased estimator is maximal among all parametric submodels," (van der Laan, 2010a). Since the Cramer-Rao lower bound corresponds with a standardized L_2 norm of $d\Psi\big(Q_n(\epsilon)\big)/d\epsilon$ evaluated at $\epsilon = 0$, this is equivalent to selecting the parametric submodel for which this derivative is maximal w.r.t. this L_2 norm.

We also seek an (asymptotically) efficient estimator. This too is achieved with the above described fluctuated update $Q_n(\epsilon)$ because the score of our parametric submodel at zero fluctuation equals the efficient influence curve of the pathwise derivative of the target parameter, Ψ (also evaluated at $\epsilon = 0$).

TMLE thus essentially consists in 1) selecting a submodel $Q_g(\epsilon)$ possibly indexed by nuisance parameter g, and 2) a valid loss function $L(Q, O) : (Q, O) \rightarrow L(Q, O) \in \mathbb{R}$. Given these two elements, TMLE solves

$$P_n \left\{ \frac{d}{d(\epsilon)} \Big(L\big(Q_n^*(\epsilon)\big) \Big)_{\epsilon=0} \right\} = 0, \tag{1.6}$$

so if this "score" is equal to the efficient influence curve, $D^*(Q_n^*, g_n)$, then we have that Q_n^* solves $P_n D^*(Q_n^*, g_n) = 0$. Now a result from semi-parametric theory is that solving this efficient score for the target parameter yields, under regularity conditions (including the requirement that Q_n and g_n consistently estimate Q_0 and g_0, respectively), an asymptotically linear estimator with influence curve equal to $D^*(Q_0, g_0)$. The TMLE of the target parameter is therefore efficient. Moreover, the TMLE is double-robust in that it is a consistent estimator of $\Psi(Q_0)$ if either Q_n or g_n is consistent.

TMLE acquires this property by choosing the fluctuation function, Q^*, such that it includes a term derived from the efficient influence curve of $\Psi(Q_0)$.

The following theorem presents the efficient influence curve for a parameter like

the ones described above. The content of the theorem will make it immediately apparent why the fluctuation function described subsequently takes the form it does; i.e., it will be seen how the terms in the efficient influence curve lead directly to the form of the fluctuation function, $Q_{L(j)n}(\epsilon)$.

1.3.2 Efficient Influence Curve

We repeat here Theorem 1 from van der Laan (2010a).

Theorem 1 *The efficient influence curve for $\Psi(Q_0) = E_0 Y_d$ at the true distribution P_0 of O can be represented as*

$$D^* = \Pi(D_{IPCW} \mid T_Q),$$

where

$$D_{IPCW}(O) = \frac{I\big(\bar{A} = d(\bar{L})\big)}{g\big(\bar{A} = d(\bar{L}) \mid X\big)} Y - \psi.$$

T_Q is the tangent space of Q in the nonparametric model, X is the full data (in the present context the full data X would be defined as $\{N(k) : k = 0, 1, 2, ..., K\}$) and Π denotes the projection operator onto T_Q in the Hilbert space $L_0^2(P_0)$ of square P_0-integrable functions of O, endowed with inner product $\langle h_1, h_2 \rangle = E_{P_0} h_1 h_2(O)$.

This subspace

$$T_Q = \sum_{j=0}^{2} T_{Q_{L(j)}}$$

is the orthogonal sum of the tangent spaces $T_{Q_{L(j)}}$ of the $Q_{L(j)}$-factors, which consists of functions of $L(j), Pa\big(L(j)\big)$ with conditional mean zero, given the parents $Pa\big(L(j)\big)$ of $L(j), j = 0, 1, 2$. Recall also that we denote $L(2)$ by 'Y.'

Let

$$D_j^*(Q, g) = \Pi\big(D_j \mid T_{Q_{L(j)}}\big).$$

Then

$$D_0^* = E\big(Y_d \mid L(0)\big) - \psi,$$

$$D_1^* = \frac{I[A(0) = d_0(L(0))]}{g[A(0) = d_0(L(0)) \mid X]} \times \big\{C_{L(1)}(Q_0)(1) - C_{L(1)}(Q_0)(0)\big\}\big\{L(1) - E\big(L(1) \mid L(0), A(0)\big)\big\},$$

$$D_2^* = \frac{I[\bar{A} = d(\bar{L})]}{g[\bar{A} = d(\bar{L}) \mid X]}\big\{L(2) - E\big(L(2) \mid \bar{L}(1), \bar{A}(2)\big)\big\},$$

where, for $\delta = \{0, 1\}$ *we used the notation*

$$C_{L(1)}(Q_0)(\delta) \equiv E\big(Y_d \mid L(0), A(0) = d(L(0)), L(1) = \delta\big).$$

We note that

$$E\big(Y_d \mid L(0), A(0) = d_0(L(0)), L(1)\big) = E\big(Y \mid \bar{L}(1), \bar{A} = d(\bar{L})\big).$$

We omit the rest of the theorem as presented in van der Laan (2010a) as it pertains to data structures with up to T time points, $T \in \mathbb{N}$.

As mentioned above, TMLE solves the efficient influence curve equation, $P_n D^*(Q_n^*, g_n) = 0$. This is accomplished by adding a covariate to an initial estimator $Q_{L(j)}^{(0)}$ as follows. (Here $L(j)$ is taken as binary.)

$$logit\big(Q_{L(j)n}(\epsilon)\big) = logit\Big(Q_{L(j)n}^{(0)}\Big) + \epsilon C_{L(j)}(Q_n, g_n), \tag{1.7}$$

where, for example,

$$C_{L(1)}(Q, g) \equiv \frac{I[A(0) = d_0(L(0))]}{g[A(0) = d_0(L(0)) \mid X]}\big\{C_{L(1)}(Q_0)(1) - C_{L(1)}(Q_0)(0)\big\},$$

with $C_{L(1)}(Q_0)(\delta)$ as defined in Theorem 1, and

$$C_{L(2)}(Q, g) \equiv \frac{I\big(\bar{A} = d(\bar{L})\big)}{g\big(\bar{A} = d(\bar{L}) \mid X\big)}.$$

It immediately follows that this choice of $Q_{L(j)}(\epsilon)$ yields a score that is equal to the efficient influence curve at $\epsilon = 0$ as claimed.

1.3.3 Implementation of the TMLE's

Below we briefly describe two different procedures for the fitting of ϵ, which we call the *one-step* and *iterative* approaches, which result in two distinct targeted maximum likelihood estimators. The iterative approach estimates a common ϵ for all factors for which a fluctuation function is applied, and the one-step estimator fits each factor separately. In the latter case 'ϵ' in equation (1.7) should be replaced with 'ϵ_j.'

There is at least one other method of fitting ϵ that we are aware of, which we describe in the next chapter. The idea here is to start with an initial estimator $Q_n(\epsilon)$, where this initial estimator is defined as in equation (1.7), with ϵ chosen at some initial value (say $-1 \leq \epsilon \leq 1$). This estimator is then plugged into the empirical efficient influence curve estimating equation, and then numerical analysis methods are used to find

$$\epsilon_n = \underset{\epsilon}{argmin} \, |P_n D^*(Q_n(\epsilon), g_n)|,$$

where g_n is an estimate of the treatment mechanism, which can be either given or estimated from the data, and $\epsilon \in [a, b]$ where a, b are assumed to bracket the solution ϵ_n. We describe this procedure in detail in chapter 2.

It's worth noting that the number of different TMLE's is not limited to the number of methods for fitting the fluctuation function. Targeted maximum likelihood estimators can also be indexed by different initial estimators, $Q^{(0)}$. Thus, for example, one may choose an initial estimator corresponding to a parametric model for Q_0, or, as we prefer, choose one corresponding to a data-adaptive estimator. The latter can be partitioned into many varieties as well; thus the number of initial estimators is vast, and this translates to a corresponding number of possible TMLE's. We explore this flexibility in some detail in chapter 3. The class of TMLE's is thus defined by the fact that they all apply a specific fluctuation function to the initial estimator $Q^{(0)}$ (which is explicitly designed so that the derivative of the loss function at zero fluctuation is equal to the efficient influence curve), independent of the choice of $Q^{(0)}$, and a loss function for the purposes of estimating ϵ.

Of course, some choices for $Q^{(0)}$ are better than others in that they will be better approximations of Q_0. Doing a good job on the initial estimator has important performance consequences, which is one good reason to pursue an aggressive data-adaptive approach.

One-Step TMLE

The one-step TMLE exploits the fact that estimates of the conditional distributions of Y and Y_d are not required in order to compute the clever covariate term of $Q_{L(2)}(\epsilon)$, the latter being the final Q_0 term in the time-ordering of the factors (for a two-stage sequential randomized trial). This allows one to update $Q^{(0)}_{L_d(2)} \equiv P\big(Y_d = 1 \mid L_d(1), L(0)\big) = E_{Q^{(0)}}\big(Y_d \mid L_d(1), L(0)\big)$ with its fluctuation $\epsilon_2 C_{L(2)}(Q, g)$ first, then use this updated (i.e., fluctuated) estimate $Q^*_{L(2)}$ in the updating step of the $Q_{L(1)}$ term. We remind the reader that the efficient influence curve—and hence $C_{L(j)}(Q, g)$—is parameter-specific, and therefore different parameters (which in our context amounts to different EY_d indexed by d) will have different realizations of the clever covariates.

As with the maximum likelihood estimator (discussed in section 2.2), both estimators (one-step and iterative) require an initial estimate $Q^{(0)}_{L(j)}$ of $Q_{L(j)}$ for $j = 0, 1, 2$, where $Q^{(0)}_{L(0)} \equiv P_{Q^{(0)}}(L(0))$ will just be estimated by the empirical distribution of $L(0)$. Thus the estimates $Q^{(0)}_{L(j)}$, $j = 1, 2$ would just be, e.g., the ML estimates if that is how one obtains one's initial estimate of Q_0. (However, as mentioned previously, we strongly recommend a data-adaptive/machine learning approach for obtaining the initial estimators.) Upon obtaining these initial estimates of Q_0, one then computes an "updated" estimate $Q^*_{L(2)}$ by fitting the coefficient ϵ_2 using (in this case of binary factors), logistic regression. The estimate of ϵ_2 is thus an MLE. This means computing a column of values of $C_{L(2)}$ (one value per observation) and then regressing the outcome $L(2)$ on this variable using the logit of the initial prediction (based on $Q^{(0)}_{L(2)}$) as offset. That is, for each observation a predicted value of $L(2)$ on the logit scale is generated based on the previously obtained $Q^{(0)}_{L(2)}$. Then $\epsilon_{2,n}$ is found by regressing $L(2)$ on the computed column $C_{L(2)}$ with $logit\left(Q^{(0)}_{L(2)}\right)$ as offset. (This is achieved in R with the offset argument in the *glm* function.)

Note that this clever covariate, $C_{L(2)}$, requires an estimate of $g\big(\bar{A} \mid X\big) = g(\bar{A} \mid L(0), L(1)\big)$ (the latter equality valid under the sequential randomization assumption). With $A(0)$ random and $A(1)$ a function of $L(1)$ only, and if $L(1)$ is binary or discrete, this estimate is easily obtained non-parametrically. If $L(1)$ is continuous, some modeling will be required.

Having obtained an estimate $Q^*_{L(2)}$ (which is parameter-dependent, and hence *targeted* at the parameter of interest), one then proceeds to update the estimate of $Q_{L(1)}$ by fitting the coefficient $\epsilon_{1,n}$—again using logistic regression if $L(1)$ is binary. Note that the clever covariate $C_{L(1)}(Q, g)$ involves an estimate of $Q_{L(2)}$. Naturally, we use our best (parameter-targeted) estimate for this, $Q^*_{L(2)}$, which was obtained in the previous step. $Q^* \equiv \big(Q^*_{L(1)}, Q^*_{L(2)}\big)$ now solves the efficient influence curve equation, and iterating the above procedure will not result in an updated estimate of Q^*—i.e., the estimates of ϵ will be zero if the

procedure is repeated using the Q^* obtained in the previous round as initial estimator. Armed now with the updated estimate Q^*, we obtain the one-step TMLE, $\Psi(Q^*)$, from the G-computation formula (1.5) for our parameter of interest with Q^* in place of Q_0.

When $L(1)$ is multilevel—say, four levels—one can model $Q_{L(1)}$ as follows. Code each of the categories for $L(1) \in \{0, 1, 2, 3\}$ as a binary indicator variable, $L(1, m),\ m = 0, 1, 2, 3$:

$$P\Big(L(1) = m \,|Pa\big(L(1)\big)\Big) \tag{1.8}$$

$$= P\Big(L(1) = m \mid L(1) \geq m, Pa\big(L(1)\big)\Big) P\Big(L(1) \geq m \mid Pa(L(1))\Big) \tag{1.9}$$

$$= P\Big(L(1, m) = 1 \mid L(1) \geq m, Pa\big(L(1)\big)\Big) \times \tag{1.10}$$

$$\prod_{m'=0}^{m-1} \Big\{1 - P\Big(L(1, m') = 1 \mid L(1) \geq m', Pa\big(L(1)\big)\Big)\Big\} \tag{1.11}$$

$$= Q_{L(1,m)}\Big(1, \bar{L}(1, m-1) = 0, Pa\big(L(1)\big)\Big) \times \tag{1.12}$$

$$\prod_{m'=0}^{m-1} Q_{L(1,m')}\Big(0, \bar{L}(1, m'-1) = 0, Pa\big(L(1)\big)\Big), \tag{1.13}$$

where $\bar{L}(1, s) = \big(L(1, s), L(1, s-1), ..., L(1, 0)\big)$. In this way, the conditional density of each binary factor of $L(1)$, $Q_{L(1,m)}$, can be estimated using logistic regression. We now denote $Q_{L(1)} = \prod_{m=0}^{3} Q_{L(1,m)}$.

To estimate these binary conditional densities, one creates a new data set analogous to a repeated measures data set, in which the number of rows corresponding to each observation is determined by the value of m for which $L(1, m) = 1$. For example, suppose that for individual i, $L(1)_i = 2$ and therefore $L(1, 2)_i = 1$. Then i will contribute three rows of data where the values in the cells for each row are identical except for two columns: a column that denotes an indicator and an adjacent column corresponding to the increasing values of m from 0 to 2. The rows for the indicator column for this individual are 0 up until $m = 2$ (at which the indicator is 1), and the next row is the first row for the next individual in the dataset. One now performs a logistic regression of the column corresponding to the indicator on the parents of $L(1)$, including the column for m.

Now with conditional densities for these binary indicator variables in hand, one can proceed with the targeting step. Each $Q_{L(1,m)},\ m = 0, 1, 2, 3$ is updated by adding a clever covariate term. The terms are again derived from the corresponding part of the efficient influence curve associated with the likelihood of the data, as factorized according to this new data structure with binary indicator variables (see Appendix A). One can see from these terms that the

updating proceeds as above for the binary $L(1)$ case, i.e., one computes $C_{L(2)}$ first, then the terms $C_{L(1,m)}$, $m = 0, 1, 2, 3$ in sequence backwards in time, starting with $C_{L(1,3)}$, and performs logistic regression to obtain the estimates of ϵ. Again, this process of computing the clever covariates and estimating the corresponding $\epsilon's$ converges in one round.

Iterative TMLE

The procedure here corresponds to estimating ϵ with the MLE,

$$\epsilon_n = \underset{\epsilon}{argmax} \prod_{j=1}^{2} \prod_{i=1}^{n} Q_{L(j),n}(\epsilon)(O_i).$$

In contrast to the one-step approach, here we estimate a single/common ϵ for all factors $Q_{L(j)}$, $j = 1, 2$.

This iterative approach requires treating the observations as repeated measures. Thus, (assuming $L(1)$ binary for the moment), each observation contributes two rows of data, and instead of a separate column for $L(1)$ and $L(2)$, the values from these columns are alternated in a single column one might call "outcome." Thus the first two rows in the data set correspond to the first observation. Both rows are the same for this first observation except for three columns: those for outcome, offset and clever covariate. There are no longer separate columns for $L(1)$ and $L(2)$, nor for the offsets, and there is likewise a single column for $C_{L(j)}$. The rows for all three columns alternate values corresponding to $j = 1$ and $j = 2$ (as described for $L(j)$).

If $L(1)$ is multi-level, the repeated measures for each observation consists of the rows described in the previous section, plus one row for $L(2)$.

Maximum likelihood estimation of ϵ is then carried out by running logistic regression on the outcome with $C_{L(j)}$ as the sole covariate, and with the logit of the initial estimator, $logit\left(Q^{(0)}_{L(j)}\right)$, as offset. This value of ϵ_n is used as coefficient for the clever covariates in the $Q_{L(j)}(\epsilon)$ terms for the next iteration. Note that $C_{L(1)} = C_{L(1)}(Q_n, g_n)$. Thus for the k^{th} iteration ($k = 1, 2, ...$), $C^{(k)}_{L(1)} = C^{(k)}_{L(1)}\left(Q^{(k-1)}_n, g_n\right)$, and g_n is not updated. The process can be iterated till convergence. Convergence is hardly required, however, if the difference $|\psi^{(k-1)}_n - \psi^{(k)}_n|$ is much smaller than $var\left(\psi^{(k-1)}_n\right)$. Here $\psi^{(k)}_n \equiv \Psi\left(Q^{(k)}(\epsilon)\right)$ is the k^{th} iteration TMLE of the parameter, and the estimated variance, $var_n\left(\psi^{(k-1)}_n\right)$ can be used in place of the true variance. Our simulations suggest that the iterated values of $\psi^{(k)}_n$ are approximately monotonic, and in any case, the value of $|\epsilon_n|$ for successive iterations typically diminishes more than

an order of magnitude. The latter fact implies that successive iterations always produce increasingly smaller values of the absolute difference $|\psi_n^{(k-1)} - \psi_n^{(k)}|$, which means that once this difference meets the above stated criterion, the process is complete for all practical purposes.

1.4 Simulations

We simulated data corresponding to the data structure described in section 1.2 under varying conditions. The specifics of the data generation process are given in Appendix B. The conditions chosen illustrate the double-robustness property of TMLE and EE, and behavior at various sample sizes. We report on simulations in which $A(0)$ was assigned randomly but $A(1)$ was assigned in response to an individual's $L(1)$; the latter corresponding to an individual's intermediate response to treatment $A(0)$. The specification of these dynamic regimes are given in the following section.

Simulations were divided into two main cases: binary $L(1)$, and discrete $L(1)$ with four levels. For each simulated data set, we computed the estimate of the target parameter $\Psi(P_0) \equiv EY_d$ for three specific rules using the following estimators: 1) One-step TMLE; 2) Iterative TMLE; 3) Inverse Probability of Treatment Weighting (IPTW); 4) Efficient Influence Curve Estimating Equation Methodology (EE); 5) Maximum Likelihood Estimation using the G-computation formula. In the *Results* subsection we give bias, variance and relative MSE estimates.

Here is a brief description of each of the estimators examined.

- *Maximum Likelihood*
  The (parametric) MLE requires a parametric specification of $Q_{L(j)}$ for computation of the parameter estimate, $\Psi(Q_0)$. The form used (e.g., $Q_{L(j),n} = expit[m(\bar{L}(j-1), \bar{A}(j-1) \mid \beta_n)]$ for some function $m(\cdot \mid \cdot)$) was either that of the correct $Q_{L(j)}$ or a purposely misspecified form, and in either case the MLE of the coefficients β were obtained with common software (namely, the *glm* function in the R language). The estimate of EY_d was then computed using the G-computation formula (1.5), which, e.g., with binary Y and binary $L(1)$, and using the empirical distribution of $L(0)$ is

$$\begin{aligned}\Psi(Q_0) =& \frac{1}{n}\sum_{i=1}^{n}\sum_{y} y \sum_{l(1)} Q^d_{L(1)}(L(0)_i, l(1)) Q^d_{L(2)}(L(0)_i, l(1), y) \\ =& \frac{1}{n}\sum_{i=1}^{n} \Big\{ Q^d_{L(1)}(L(0)_i, l(1)=1) Q^d_{L(2)}(L(0)_i, l(1)=1, y=1) \\ & + Q^d_{L(1)}(L(0)_i, l(1)=0) Q^d_{L(2)}(L(0)_i, l(1)=0, y=1) \Big\}.\end{aligned}$$

The maximum likelihood estimator, which is a substitution estimator, can thus be expressed as $\Psi_n^{MLE} = \Psi\left(Q^{(0)}\right)$, where for each factor $Q^d_{L(j)}$ in the G-computation formula, the corresponding MLE, $Q^{(0)d}_{L(j)}$ is substituted, and where $Q^{(0),d} \equiv Q^{MLE,d}$.

The estimator thus requires estimations of $Q_{L(j)} \equiv P(L(j) \mid Pa(L(j)))$, which as mentioned above, were correctly specified for one set of simulations and incorrectly specified for another.

- *One-Step TMLE*
  See Implementation section above. The initial estimator of Q_0 is the MLE estimator given above.

- *Iterative TMLE*
  See Implementation section above. Here also the initial estimator of Q_0 is the MLE estimator.

- *IPTW*
  The IPTW estimator is defined to be

$$\psi_n^{IPTW} = \frac{1}{n}\sum_{i=1}^{n} Y_i \frac{I(\bar{A}_i = d(\bar{L})}{g[\bar{A}_i = d(\bar{L}) \mid X_i]}.$$

  As with TMLE, this estimator requires estimation of $g[\bar{A} = d(\bar{L}) \mid X]$, which for binary factors and binary treatment is a straightforward non-parametric computation. The IPTW estimator is known to become unstable when there are ETA violations, or practical ETA violations. Adjustments to the estimator that compensate for these issues have been proposed (Bembom and van der Laan, 2008). In the simulations at hand, $g[\bar{A} = d(\bar{L}) \mid \bar{L}]$ was bounded well away from 0 and 1 but was nevertheless not estimated at all (the true distribution of $A \mid X$ was used). However, van der Laan and Robins (2003) show that there is some efficiency gain in estimating $g(\bar{A} \mid \bar{L})$ over using the known true g.

- *Estimating Equation Method*
  This method solves the efficient influence curve estimating equation in ψ. That is,

$$\psi_n^{EE} = P_n E_{Q_n}(Y_d \mid L(0)) + \frac{1}{n}\sum_i \left\{ D_{1,n}^*(O_i) + D_{2,n}^*(O_i) \right\},$$

with $D_{1,n}^*, D_{2,n}^*$ as given in Theorem 1 except that the true conditional expectations of Y and of Y_d in the expressions for D_1^* and D_2^* are replaced with their respective sample estimates. The only difference between this estimator and the so-called augmented inverse probability of censoring weights (A-IPCW) estimator is in the way the expression for the efficient influence curve is derived. The results for the A-IPCW estimator should be identical to those for the one we describe here.

Just as with the TMLE, this estimator requires model specifications of $Q_{L(j)}$, $j = 1, 2$ for estimation of $E(Y_d \mid L(0))$ and for the elements of D_1^*, D_2^* that involve conditional expectations of Y_d and of Y. Here again we used the ML estimates of $Q_{L(j)}$, under both correct and incorrect model specification scenarios, i.e., we used $Q_n = Q^{(0)}$ for the factors involving estimates of Q_0 in the estimating equation above. (See description of the Maximum Likelihood Estimator above.)

1.4.1 Some Specific Treatment Rules

We considered several treatment rules, one set for binary $L(1)$ (three different rules), and a necessarily different set (also three separate rules) for the discrete $L(1)$ case. This permits easy computation of the natural parameters of interest $EY_{d_i} - EY_{d_j}$, for $i \neq j$, where in our case, $i, j = 1, 2, 3$. Indeed such parameters are arguably the ultimate parameters of interest to researchers utilizing longitudinal data of the type described here, since they implicitly give the optimum treatment rule among those considered. As the number of discrete levels of $L(1)$ increases, one can begin considering indexing treatment rules by threshold levels θ of $L(1)$ such that, e.g., assuming binary $A(0)$ and $A(1)$, one could set A(1) according to $A(1) = [1 - A(0)]I(l(1) < \theta) + [A(0)]I(l(1) \geq \theta)$.

Binary L(1)

In the binary $L(1)$ case, we considered the following three treatment rules

- *Rule 1*. $A(0) = 1$, $A(1) = A(0) * I(L(1) = 1) + (1 - A(0)) * I(L(1) = 0)$. In words, set treatment at time 0 to treatment 1, and if the patient does well on that treatment as defined by $L(1) = 1$, continue with same treatment at time 1. Otherwise, switch at $A(1)$ to treatment 0.
- *Rule 2*. $A(0)$ either 0 or 1, and $A(1) = A(0)$. That is, $A(0)$ can be either 0 or 1, but whatever it is, stay on the same treatment at time 1, independent of patient's response to treatment $A(0)$.

- *Rule 3*. $A(0) = 0$, $A(1) = A(0) * I(L(1) = 1) + (1 - A(0)) * I(L(1) = 0)$. In words, set treatment at time 0 to 0 and if the patient does well, stay on treatment 0 at time 1, otherwise switch to treatment 1 at $A(1)$. This is identical to Rule 1 except that patients start on treatment 0 instead of treatment 1.

Note that estimation of, or evaluation of, a rule-specific parameter does not require that patients were actually assigned treatment in that manner, i.e., according to the rule. If patients were assigned treatment randomly, then one simply needs to know which individuals in fact followed the rule in order to estimate the rule-specific mean outcome. In this case, and with $P(A(0) = 1) = P(A(1) = 1) = 0.5$, one could also construct the simple, consistent estimator $(1/n_d) \sum_i Y_i I(\bar{A}_i = d(\bar{L}_i))$, where $n_d = \sum_i (\bar{A}_i = d(\bar{L}_i))$, but this estimator is inefficient relative to the double-robust estimators.

On the other hand, if treatment was indeed assigned according to, e.g., one of the above treatment rules, then $L(1)$ is a time-dependent confounder. These are really the cases of interest. If one's estimator does not adjust for confounding in these cases it will be biased. All the estimators we compared attempt to adjust for confounding in one way or another.

Discrete L(1) with Four Values

With discrete-valued $L(1)$ ($L(1) \in \{0, 1, 2, 3\}$), the treatment rules were necessarily modified slightly to accommodate the additional values. The analog of rule 1 above, for example, is of the form

- $A(0) = 1$, $A(1) = A(0) * I(L(1) > l(1)) + (1 - A(0)) * I(L(1) \leq l(1))$ for some $l(1) \in \{0, 1, 2, 3\}$.

1.4.2 Simulation Results

Notes on the tables

Estimates of bias, variance and relative mean squared error (Rel MSE) are presented for the TMLEs and several comparison estimators. We define relative MSE for each estimator as the ratio of its MSE to that of an efficient, unbiased estimator. The efficiency bound here is the variance of the efficient influence curve. Thus for each estimator ψ_n of ψ_0,

$$\text{Rel MSE} \equiv \frac{(E(\psi_n) - \psi_0)^2 + var(\psi_n)}{var(D^*(Q, g))/n}, \tag{1.14}$$

where D^* is the efficient influence curve for the relevant parameter, Ψ^F. In fact, the value used in these computations for $var\left(D^*(Q,g)\right)$ is itself an estimate computed from taking the variance of $D^*(Q_0,g_0)(O)$ from a large number of observations generated from P_0.

The bias values shown are not accurate to much less than 10^{-3}. This is because the true parameter values were also obtained by simulation from the true P_d for each rule d with a large number of observations. Thus bias estimates that appear to be smaller than this should be viewed as simply being $< 10^{-3}$. We indicate these estimates with an asterisk.

Qm, gc denotes results where g (the treatment mechanism) was correctly specified, but $Q_{L(2)}$ was purposely misspecified. Qc, gc are simulations for which both Q and g are correctly specified. In an SRCT, we expect g to be known and thus did not perform analyses with a misspecified g. For each trial scenario we present results for both Qc, gc and Qm, gc. Note that the IPTW estimator is not affected by whether or not Q_n is correctly specified, since it does not estimate Q_0.

Varying numbers of simulations were done under the different scenarios. The number of simulations under each configuration (i.e., a given scenario and either Qc, gc or Qm, gc) ranged from 1990 to 5000 depending on computation time.

Confidence Intervals and Coverage Estimates

Table 1.3 gives influence curve-based estimates of the true coverage for computed 95% confidence intervals for the two TMLEs. The latter were computed for each simulated data set by estimating the variance of the efficient influence curve using that data set.

Scenario I: Binary L(1); A(1) Assigned in Response to L(1)

For brevity we only include the performance of the estimators for a single parameter, EY_1. The results for the other treatment-rule-specific parameters are similar.

Scenario II: Discrete L(1); A(1) Assigned in Response to L(1)

With discrete $L(1)$ we modeled the binary factors $Q_{L(1,m)}$ similarly to the way these factors were generated, i.e., using a hazard approach (see Appendix B). Thus each binary factor is modeled with logistic regression: as with the binary case, an initial estimate $Q^{(0)}_{L(1,m)}$ is obtained by logistic regression (where

Qc, gc

		TMLE (1-step)	TMLE (Iter)	IPTW	MLE	EE
$n = 100$	Bias	3.0e-3	2.8e-3	-1.5e.3	1.2e-3	1.8e-3
	Var	3.9e-3	3.9e-3	1.1e-2	3.9e-3	3.8e-3
	Rel MSE	1.3	1.3	3.9	1.3	1.3
$n = 250$	Bias	*	*	-2.4e-3	1.0e-3	*
	Var	1.3e-3	1.3e-3	4.6e-3	1.3e-3	1.3e-3
	Rel MSE	1.1	1.1	3.9	1.1	1.1
$n = 500$	Bias	*	*	-1.7e-3	*	*
	Var	6.3e-4	6.3e-4	2.3e-3	6.3e-4	6.3e-4
	Rel MSE	1.1	1.1	4.0	1.1	1.1

Qm, gc

		TMLE (1-step)	TMLE (Iter)	IPTW	MLE	EE
$n = 100$	Bias	3.9e-3	3.5e-3	1.5e-3	-1.2e-1	-1.2e-3
	Var	4.5e-3	4.5e-3	1.1e-2	2.8e-3	4.1e-3
	Rel MSE	1.6	1.5	3.9	6.3	1.4
$n = 250$	Bias	1.4e-3	1.1e-3	-2.4e-3	-1.3e-1	-1.3e-3
	Var	1.7e-3	1.7e-3	4.6e-3	1.1e-3	1.6e-3
	Rel MSE	1.4	1.4	3.9	14.6	1.4
$n = 500$	Bias	*	*	-1.7e-3	-1.3e-1	*
	Var	8.7e-4	8.6e-4	2.3e-3	5.7e-4	8.3e-4
	Rel MSE	1.5	1.5	4.0	28.5	1.4

Table 1.1: Scenario I Results: Performance of the various estimators in estimating EY_1 at various sample sizes. 'Qc, gc': Q correctly specified, g correctly specified; 'Qm, gc': Q misspecified, g correctly specified. Iterative TMLE estimates in this table were for the 5th iteration. Asterisks indicate bias $<$ 10e-3.

Qc, gc

		TMLE (1-step)	TMLE (Iter)	IPTW	MLE	EE
$n = 100$	Bias	-3.1e-3	-3.0e-3	-3.2e-3	-2.6e-3	-3.3e-3
	Var	5.5e-3	5.5e-3	2.0e-2	4.9e-3	5.4e-3
	Rel MSE	1.1	1.1	4.0	1.0	1.1
$n = 200$	Bias	-1.5e-3	-1.4e-3	3.8e-3	1.2e-3	-1.5e-3
	Var	2.6e-3	2.6e-3	1.2e-2	2.3e-3	2.6e-3
	Rel MSE	1.0	1.0	4.1	0.9	1.0
$n = 500$	Bias	*	*	1.3e-3	*	*
	Var	1.0e-3	1.0e-3	4.3e-3	9.0e-4	1.0e-3
	Rel MSE	1.0	1.0	4.2	0.9	1.0

Qm, gc

		TMLE (1-step)	TMLE (Iter)	IPTW	MLE	EE
$n = 100$	Bias	-1.7e-3	-1.7e-3	-3.2e-3	-7.0e-2	-3.2e-3
	Var	5.2e-3	5.2e-3	2.0e-2	2.9e-3	5.1e-3
	Rel MSE	1.0	1.0	4.0	1.5	1.0
$n = 200$	Bias	-1.9e-3	-1.9e-3	3.8e-3	-7.0e-2	-2.2e-3
	Var	2.6e-3	2.6e-3	1.2e-2	1.5e-3	2.6e-3
	Rel MSE	1.0	1.0	4.1	2.5	1.0
$n = 500$	Bias	*	*	1.3e-3	-7.0e-2	*
	Var	1.1e-3	1.1e-3	4.3e-3	6.4e-4	1.1e-3
	Rel MSE	1.1	1.1	4.2	5.5	1.0

Table 1.2: Scenario II Results: Performance of the various estimators in estimating a single parameter, EY_1, for various sample sizes. 'Qc, gc' means Q correctly specified, g correctly specified, while 'Qm' means Q misspecified. Iterative TMLE estimates in this table were for the 3rd iteration. Asterisks indicate bias $<$ 10e-3.

this estimator could be correctly or incorrectly specified) and a corresponding fluctuation function applied.

Small Sample Results

We also simulated data under scenario II above for a sample size of 30. We anticipated efficiency differences (if any) between the iterative and one-step TMLEs would show up at this very small sample size (see section 1.4.3). We saw no significant difference in the variance of these two estimators, however. The performance of the TMLEs at this sample size is remarkable, particularly under model misspecification, and we felt these results warranted a separate table (see Table 1.4).

Scenario I

	n = 100	250	500
Qc,gc			
TMLE (1-step)	0.85	0.92	0.93
TMLE (iter)	0.85	0.92	0.94
Qm,gc			
TMLE (1-step)	0.91	0.93	0.94
TMLE (iter)	0.91	0.93	0.94

Scenario II

	n = 100	200	500
Qc,gc			
TMLE (1-step)	0.88	0.91	0.94
TMLE (iter)	0.89	0.91	0.94
Qm,gc			
TMLE (1-step)	0.91	0.92	0.95
TMLE (iter)	0.91	0.92	0.95

Table 1.3: Coverage for nominal 95% confidence intervals under both data generation scenarios for the two TMLEs at various sample sizes.

Qc,gc

	Bias	Var	Rel MSE
TMLE (1-step)	-0.016	0.023	1.4
TMLE (iter)	-0.021	0.022	1.4
IPTW	3.7e-3	0.069	4.1
MLE	-0.035	0.021	1.3
EE	-0.027	0.021	1.3

Qm,gc

	Bias	Var	Rel MSE
TMLE (1-step)	-6.5e-3	0.019	1.2
TMLE (iter)	-7.0e-3	0.019	1.1
IPTW	3.7e-3	0.069	4.1
MLE	-3.0e-1	0.070	9.4
EE	-9.8e-3	0.027	1.6

Table 1.4: Scenario II Data, at $n = 30$: Performance of the various estimators in estimating EY_1. 'Qc, gc' means Q correctly specified, g correctly specified, while 'Qm' means Q misspecified. Iterative TMLE estimates in this table were for the 4th iteration.

1.4.3 Discussion

Relative efficiency for the ML estimator is almost always ≤ 1. The semi-parametric efficiency bound does not apply in general to that of an estimator based on a parametric model. Even so, when Q is correctly specified, the variance of the ML estimator appears to be very close to the semi-parametric efficiency bound when $n \geq 200$.

Of particular note is that the TMLE, EE and MLE estimators are already very close to the efficiency bound at $n = 250$ under Qc in the binary $L(1)$ case. Further, the reduction in bias in going to $n = 500$ is small in absolute terms.

Even more noteworthy is the performance of the TMLEs at the small sample size of 30 for the scenario II simulations (discrete $L(1)$). Bias and variance of both estimators are *better* when $Q^{(0)}$ is misspecified. Misspecification in this case consisted in setting $Logit(Q_{L(2)}) = 3 * L(1)$ (compare with the true data generating function given in Appendix B), but using correct specification for $Q_{L(1)}$. With $Q^{(0)}$ misspecified, the bias of both TMLEs is quite small and the variance is very close to the efficiency bound. EE also shows lower bias under incorrect Q, but not lower variance. The better performance under misspecification can be understood by noting that under correct model specification, many more parameters of the model must be fit. We expect that asymptotically, there is a gain in efficiency of the TMLEs and EE if $Q^{(0)}$ is consistently estimated, but these simulations show that a parsimonious model as initial estimator, even if misspecified, can have distinct advantages in TMLE at small sample sizes.

The effect is still noticeable at sample size 100 in the discrete $L(1)$ case. There we also see lower bias of the TMLEs under incorrect model specification than under correct model specification. This phenomenon is not present in the scenario I simulations however.

The advantage of the TMLEs' being substitution estimators also becomes apparent in these small sample results: at $n = 30$, many times the EE and IPTW estimators gave estimates outside the bounds of the model ($EY_d \in [0, 1]$). Indeed, under Qm, the EE estimator gave estimates of $EY_1 > 1$ more than 13% of the time. For more extensive performance comparisons between TMLE and other double robust estimators (including the A-IPCW estimator) under various conditions, including sparsity/positivity violation conditions, see, e.g., Porter et al. (2011), Stitelman et al. (2011), Gruber and van der Laan (2010), Stitelman and van der Laan (2010) and van der Laan and Rose (2011).

In general, under incorrect specification of Q we do not expect any of the estimators that estimate Q_0 to be *asymptotically* efficient except for the MLE, which used a much simpler model than the true model and therefore could

easily achieve a lower variance bound. Misspecification of Q in all cases meant misspecifying $Q^{(0)}_{L(2)}$ but correctly specifying $Q_{L(1)}$. Thus under Qm, gc the MLE will be biased but the TMLE and EE estimators are double robust and therefore still asymptotically unbiased under correct specification of g. Under the scenarios simulated here g is expected to be known and we therefore omitted simulations in which g is misspecified; the latter will of course result in bias of the IPTW estimator. Scenarios in which g is not known, or not completely known are also quite plausible, however; e.g., one can easily imagine settings in which assignment of $A(0)$ and/or $A(1)$ was not done in complete accordance with a defined treatment rule. Nevertheless, even in these cases, with $A(0)$ randomized and $L(1)$ discrete or binary, non-parametric estimation of g would not be difficult. If $A(0)$ is a function of $L(0)$ then some smoothing will be required for the estimate of $g(A(0) \mid L(0))$ and model misspecification is likely to arise.

The two versions of TMLE we've implemented (one-step and iterative) typically agree in their estimate of the parameter to within 1%, and in many cases to within quite a bit less than this. For the two time-point data structure we've simulated, the one-step estimator is conceptually easier to implement than the iterative approach, and slightly faster computationally. As the number of estimated factors increases (either from having multiple time points, multiple covariates in $L(j)$, $1 < j < K$, or both), the iterative method may become the more practical programming choice.

Also noteworthy is that the one-step TMLE requires estimation of two ϵ's in the binary $L(1)$ case and four ϵ's in the discrete $L(1)$ case. For the general data structure $(L(0), A(0), ...L(K), A(K), L(K+1))$ where intermediate factor $L(j)$ has t_j levels, the number of ϵ's the one-step estimator must fit is $\sum_{j=1}^{K+1}(t_j - 1)$. In contrast, the iterative TMLE performs a fitting of ϵ that is independent of K and t_j. (Though a new round of fitting occurs for each iteration, the bulk of the fitting occurs in the first iteration.) We thus expected at least a small efficiency advantage for the iterative method, though we have not observed it in the simulations presented here, even in sample sizes as low as 30.

Comparison of the TMLE and Estimating Equation Methods

The fundamental differences between targeted maximum likelihood estimation and estimating equation-based estimation have been detailed in the seminal targeted maximum likelihood paper (van der Laan and Rubin, 2006) and elsewhere (see, e.g., van der Laan and Rose, 2011). The differences bear repeating, however, and we give a synopsis of them here.

The most essential difference is summed up in the fact that a TMLE is defined as a (particular) substitution estimator—i.e, an estimator that can be represented as $\Psi(P_n^*)$ for an estimator P_n^* in the statistical model $\mathcal{M}$—and an EE

estimator is not. This difference has important ramifications.

The EE algorithm is defined by writing the efficient influence curve, $D(P)$, as an estimating function $D(\psi, \eta)$ in terms of parameter ψ and nuisance parameter η, and solving for ψ (van der Laan and Robins, 2003). In general, being able to express $D(P)$ in such a form is not a reasonable requirement for parameters and models. In contrast the TMLE algorithm (described in section 1.3.1) does not rely on the efficient influence curve's being an estimating function.

The TMLE definition also does not rely on an estimating equation's having a unique solution, while EE is only well defined if the estimating equation has a unique solution in ψ. The existence of multiple solutions of estimating equations is a common phenomenon, just as a log-likelihood can have multiple local maxima (and thus multiple solutions for the associated score equation) even though it has a unique maximum. The TMLE P_n^* of P is not defined as a solution of the equation $0 = \sum_i D^*(P)(O_i)$ in P either (it is not even sensible to state that $P_n D(P) = 0$ has a unique solution in P, since there is a whole class of P's that solve it)—it just happens to solve the efficient score equation $0 = \sum_i D^*(P_n^*)(O_i)$ as a by-product of iteratively maximizing the likelihood (or other loss) along a least-favorable submodel.

Instead of having to deal with multiple solutions of an equation, one might well be faced with an estimating equation with no solution at all (in its parameter space); this can occur, for example, under practical violations of the positivity assumption. In the estimation problem addressed in this article, with positive probability the relevant estimating equation is only solved by a negative number, or number larger than 1. We noted this behavior of the EE estimator—i.e., giving an estimate that's not even a probability—in the Results section above.

As mentioned above, dramatic differences in finite sample performance between EE (of which A-IPCW is an example) and TMLE under practical violations of the positivity assumption have been established in many settings. The erratic behavior of EE in such cases is mainly due to its not respecting the global constraints imposed by the target parameter mapping defined on the statistical model. Such differences in behavior are not expected in a sequentially randomized trial in which the treatment mechanism is known and nicely bounded away from zero, and sample size is reasonably large, but the differences can be quite apparent in observational settings or for very small sample sizes in the sequentially randomized trial setting (as seen here).

Chapter 2

A TMLE Based on Directly Solving the Efficient Influence Curve Equation

In chapter 1 we described the implementation of two distinct targeted maximum likelihood estimation (TMLE) algorithms (the "one-step" and "iterative" procedures) for estimating specified counterfactual parameters of the underlying distribution corresponding to a particular longitudinal data structure, indexed by dynamic treatment rules. We also compared their performance to that of some well-known existing estimators. The comparative advantages amongst TMLE's involve differences in computational resources needed, and in complexity of implementation. In this chapter we present a third algorithm, which we find conceptually easier to implement than either of the foregoing methods, and whose speed is comparable to that of the iterative method.

We emphasize that there are targeted maximum likelihood estimators (plural) of a given parameter, since TMLE is a class of estimation methods that utilizes *i)* a fluctuation submodel of an initial estimator and *ii)* a loss function or other empirical criterion for fitting the submodel.

The TMLEs presented in chapter 1, as well as all other heretofore implemented TMLEs independent of data type, all solve a score equation as the means of constructing the estimator. Here we present a TMLE based on a different empirical criterion, namely, solving the empirical efficient influence curve equation directly. We have been moving toward the term "targeted minimum loss-based" rather than "targeted maximum likelihood" in describing this class of estimators, and the procedure we describe here motivates this terminological adjustment since maximizing the likelihood is not involved in the construction of the estimator.

The procedure bears a superficial similarity to that of estimating equation methodology, though our procedure solves the efficient influence curve (EIC)

equation by adjusting the amount of fluctuation of a predesignated fluctuation submodel, and does not solve it in the parameter, ψ. Like all TMLE's (and unlike estimating equation-based estimators), this TMLE is a substitution estimator, and retains the associated benefits. Results from our simulations indicate that the procedure also exhibits all of the finite sample advantages of the existing TMLE procedures, which have been described in a variety of applications (van der Laan et al., 2009), and which were also seen in chapter 1.

We refer the reader to section 1.2 for a description of the likelihood, G-formula and parameter definition.

2.1 Method

2.1.1 Existing TMLEs

In chapter 1 we explained that in the TMLE method begins by obtaining an initial estimator of Q_0; we then update this estimator with a fluctuation function that is tailored specifically to remove bias in estimating the particular parameter of interest. Naturally, this means that the fluctuation function is a function of the parameter of interest. The initial estimator, Q^0 of Q_0 can be obtained in a number of ways, but we advocate a data-adaptive approach in all cases. In any case, the TMLE methods do not require any particular estimation method for Q^0, though there are clear gains if Q^0 is close to Q_0.

Upon obtaining an initial estimate Q^0, the next step in TMLE is to apply a fluctuation function to this initial estimator that is the least favorable parametric submodel through the initial estimate, Q^0, for the parameter Ψ (van der Laan and Rubin, 2006). We signify this fluctuated update $Q_n(\epsilon)$. Since the Cramer-Rao lower bound corresponds with a standardized L_2 norm of $d\Psi(Q_n(\epsilon))/d\epsilon$ evaluated at $\epsilon = 0$, this is equivalent to selecting the parametric submodel for which this derivative is maximal w.r.t. this L_2 norm.

2.1.2 Numerical Solution TMLE

General Description

Above we mentioned that existing TMLE's solve (1.6). The method we present here involves solving instead

$$P_n D^*(Q_n(\epsilon), g) = 0 \tag{2.1}$$

in ϵ, or to the same effect, selecting ϵ_s such that

$$\epsilon_s = \underset{\epsilon}{argmin} \, |P_n D^*(Q_n(\epsilon), g)|, \tag{2.2}$$

where g is either the given, known treatment mechanism or an estimate of it, and $\epsilon \in [a, b] \subset \mathbb{R}$, which interval is assumed to contain the solution to (2.1). The general idea for this method was first suggested in van der Laan and Rubin (2006). $Q_n(\epsilon)$ takes the exact form as for the loss-based TMLE's, i.e., it uses the same parametric submodel through Q^0 (see below). What remains is to choose ϵ. If the empirical EIC is well-behaved on $\epsilon \in [a, b]$ and the solution is contained in that interval, then one should be able to find an ϵ_s such that $P_n D^*(Q_n(\epsilon_s), g_n)$ is arbitrarily close to 0, which means one has effectively found an estimator $Q_n(\epsilon_s)$ of Q_0 that solves (2.1).

Accordingly, let us define $Q_s^* \equiv Q_n(\epsilon_s)$, where ϵ_s is the solution to (2.1), or to (2.2) if a finite number of candidate solutions is considered. $\Psi(Q_s^*)$ is then the corresponding "numerical methods TMLE" of $\Psi(Q_0)$. Since this choice of Q_s^* solves $P_n D^*(Q_n(\epsilon), g) = 0$, it necessarily solves (1.6) with Q_s^* in place of $Q_n(\epsilon)$. However, since the solution ϵ_s was not arrived at via application of the loss function $L(Q, O)$ assumed in (1.6), we have no assurance that the likelihood for Q_s^* has increased relative to Q^0, the latter estimator being some initial estimate of Q_0 without fluctuation applied. That is, assuming the negative log likelihood as the loss function, we have no set of conditions that guarantees that $P_n L(Q_s^*, O) \leq P_n L(Q^0, O)$. Nevertheless, Q_s^* represents a movement along the hardest submodel from some initial Q^0, which does indeed result in an estimator $\Psi(Q_s^*)$ that is less biased than $\Psi(Q^0)$, even if in practice Q_s^* does not have a greater likelihood than Q^0, though it would be surprising if it failed to. It is nevertheless encouraging to see that $P_n L(Q_s^*) \leq P_n L(Q^0)$ in practice, where in our simulations Q^0 is the standard MLE, and this was indeed the case without fail in our simulation runs. In fact, the likelihood of the numeric solution estimator was strictly greater than that of Q^0 in all simulations.

Efficient Influence Curve and Parametric Submodel

In order to obtain a numerical solution to (2.1), one of course needs the explicit form of $D^*(Q, g)$ for the parameter being estimated. The EIC for parameter $\psi = EY_d$ when $L(1)$ and $Y \equiv L(2)$ are binary, and where d is a treatment rule is given in chapter 1, adpated from Theorem 1 of van der Laan (2010a), and the EIC for ψ when $L(1)$ is discrete-valued is given in Appendix A.

The empirical estimate of D^* (say D_n^*) for a single observation substitutes the corresponding estimates $Q_{L(j),n}$ and $Q^d_{L(j),n}$ in place of $Q_{L(j)}$ and $Q_{L_d(j)}$.

It is instructive to represent D_n^* in terms of these estimates of the Q-components of the likelihood and, by implication, in terms of ϵ, which is to be selected.

Though the D^* we present here is for the case of binary $L(1)$, our simulations for this article are for discrete-valued $L(1)$ with four levels. The EIC for the parameters we identified above are more complex in the discrete $L(1)$ case than the binary case, but the method we present here is independent of the types of variables involved. We therefore develop the method for the binary $L(1)$ case to avoid unnecessary conceptual and notational complexity.

We have

$$D_n^*(O) = D_n^*(Q_n, g_n)(O) = \sum_{j=0}^{2} D_{j,n}^*(Q_n, g_n)(O)$$

where

$$D_{0,n}^* = \sum_{l(1)} \left\{ Q_{L(2),n}^d(y=1, L(0), l(1)) Q_{L(1),n}^d(L(0), l(1)) \right\} - \psi_n,$$

$$\begin{aligned} D_{1,n}^* = & \frac{I[A(0) = d_0(L(0))]}{g[A(0) = d_0(L(0)) \mid X]} \left\{ Q_{L(2),n}^d(y=1, l(1)=1, L(0)) - \right. \\ & \left. Q_{L(2),n}^d(y=1, l(1)=0, L(0)) \right\} \times \left\{ L(1) - Q_{L(1),n}(l(1)=1, A(0), L(0)) \right\}, \end{aligned}$$

$$D_{2,n}^* = \frac{I[\bar{A} = d(\bar{L})]}{g[\bar{A} = d(\bar{L}) \mid X]} \left\{ L(2) - Q_{L(2),n}\left(y=1, \bar{L}(1), \bar{A}(1)\right) \right\}, \tag{2.3}$$

and

$$\psi_n = \widehat{EY_d} = \frac{1}{n} \sum_{i=1}^{n} \sum_{l(1)} Q_{L(2),n}^{(d)}(y=1, L(0)_i, l(1)) \prod_{j=0}^{1} Q_{L(j),n}^{(d)}(L(0)_i, l(1)).$$

and where X refers to the full data. For TMLE, the EIC gives us the form of the parametric submodel, $Q(\epsilon)$ for the conditional probability of each factor $L(j)$ that is to be estimated:

$$logit(Q_{L(j)}(\epsilon)) = logit(Q_{L(j)}^0) + \epsilon C_{L(j),n}, \tag{2.4}$$

where $Q^0_{L(j)}$ is some initial estimate of $Q_{L(j)}$ (e.g., the MLE),

$$C_{L(1),n} = \frac{I[A(0) = d_0(L(0))]}{g[A(0) = d_0(L(0)) \mid X]} \left\{ Q^d_{L(2),n}(y = 1, l(1) = 1, L(0)) - Q^d_{L(2),n}(y = 1, l(1) = 0, L(0)) \right\},$$

and

$$C_{L(2),n} = \frac{I[\bar{A} = d(\bar{L})]}{g[\bar{A} = d(\bar{L}) \mid X]}.$$

Using now fluctuation submodels $Q_{L(j)}(\epsilon)$ and $Q^d_{L(j)}(\epsilon)$ given in (2.4) for the elements $Q_{L(j),n}$ and $Q^d_{L(j),n}$, respectively, in the formula above, our method attempts to solve (2.1) or (2.2) with D^*_n in place of D^*.

2.1.3 Numerical Methods for Solving Empirical Efficient Influence Curve Equation

Though complex, (2.3) for our present purposes is nothing but a one dimensional function of ϵ. For notational convenience let us thus write $f(\epsilon) \equiv P_n D^*(Q_n(\epsilon))$. If $f(\epsilon)$ is continuous and has a unique root, then the well-known bisection and secant methods of numerical analysis (see, e.g., Faires and Burden, 2003) are promising techniques for finding the root. If, further, $f(\epsilon)$ is differentiable w.r.t ϵ on the interval over which it is being evaluated, then Newton's method is also a candidate. (Other well-known methods include the *method of false position* and *Müller's method*.)

The purpose of this chapter is primarily to present the solving of the empirical EIC equation—given a specified fluctuation submodel—via numerical techniques as a method of producing a TMLE. We thus omit technical and detailed comparisons of various numerical techniques for obtaining these solutions. For a suitably well behaved function f, the specific technique employed to find ϵ_s, though central to the actual implementation of the estimator, is of secondary importance to the overall method described here. There are certainly pros and cons of each technique, which can be assessed a priori if one knows the exact form of $f(\epsilon)$ under all applicable data sets, but one generally does not have such knowledge. The advantages associated with these techniques have to do with whether or not the algorithm is guaranteed to converge, and if it does converge, how quickly. Basic texts on the subject (e.g., Faires and Burden, 2003) give an adequate treatment of these comparisons and we refer the reader there for more detail. To re-iterate: our research interest is in the performance

of a TMLE that is produced by solving (2.1) or (2.2) in the manner explained in the previous section, and we assume that in all cases of interest there is a numerical technique adequate to the task.

Nevertheless, a brief comparison of the best known techniques for the present context is in order. We have in fact implemented both the bisection and secant methods, and have not attempted Newton's method. The appeal of Newton's method is its rate of convergence (in terms of number of iterations)—under most circumstances if it does converge it has the fastest convergence rate. This is not universally true however. Moreover, the method has the drawback that for each iteration both $f(\epsilon)$ and $f'(\epsilon)$ must be evaluated. For functions that are computationally intensive to evaluate, as in our case, this undercuts the advantage of requiring fewer iterations for a given tolerance compared to the secant method, and could even make Newton's method slower to converge in real time, even if in fewer iterations. The latter fact combined with the added complexity of implementation make the possible gains of Newton's method over the secant method negligible in our case. We thus pursued the secant method as our primary numerical method, having first implemented the bisection method.

Most worthy of mention in comparing these latter two numerical techniques is that 1) the bisection method is guaranteed to converge if the function of interest has a root on the initially specified interval, and the secant method is not (though this is not problematic in our context—see below) and 2) the bisection method is much slower to converge than the secant method in general. In our context the latter factor drives the choice between these two numerical techniques. (Recall that we seek estimators that have computational advantages in the longitudinal setting, which setting is generalizable to any number of time points, and multiple intermediate outcomes per time point.) Though we have implemented the bisection method, we found that in all cases tested, the secant method converged in far fewer iterations for a given tolerance (usually chosen to be $\sim 10^{-6}$). The difference in the values of the solution ϵ_s produced by the two methods can be made arbitrarily small by performing enough iterations. Since the secant method is superior in every way applicable to our function (except guaranteed convergence) we focus entirely on it as the chosen technique. As mentioned above, lack of guaranteed convergence is not a concern here, which we address in detail in the Discussion section. We therefore give a brief description of how to apply the secant method for our function of interest, $f(\epsilon) \equiv P_n D^*(Q_n(\epsilon))$.

Secant Method

The secant method is based on a sequence of approximations to the root of a function, generated by drawing secant lines through, in our case, the points $(\epsilon_k, f(\epsilon_k))$ and $(\epsilon_{k+1}, f(\epsilon_{k+1}))$, $k = 0, 1, ..., K$. The zero of each such line is

computed and this defines the position of the next approximation, ϵ_{k+2}. The initial values (ϵ_0, ϵ_1) need not bracket the solution though the closer they are to it, the more rapidly the algorithm will converge. Starting with initial approximations (ϵ_0, ϵ_1), the first iteration produces a new approximation

$$\epsilon_2 = \epsilon_1 - \frac{(\epsilon_1 - \epsilon_0) f(\epsilon_1)}{f(\epsilon_1) - f(\epsilon_0)}.$$

This result follows from a straight-forward application of point-slope algebra. The next iteration uses (ϵ_1, ϵ_2) as starting values and the process is iterated until $|f(\epsilon_k)| \leq T$ where T is the tolerance deemed sufficient. In our case, the difference in successive estimates $|\psi_k - \psi_{k+1}|$ was typically on the order of $|P_n D^*(\epsilon_k)|$. Thus, an ϵ_k that yields a $|P_n D^*(\epsilon_k)| \leq 10^{-3}$ is quite sufficient, though for our simulations we used $T = 10^{-6}$.

2.2 Simulations

We simulated data corresponding to the data structure described in section 1.2 for discrete-valued $L(1)$ under correct and incorrect model specification, and at various sample sizes. Incorrect model simulations were done to illustrate the double-robustness property of the TMLE's. $A(0)$ was assigned randomly but $A(1)$ was assigned in response to an individual's $L(1)$; the latter corresponding to an individual's intermediate response to treatment $A(0)$. We gave the specification of these dynamic regimes in section 1.4.1.

For each simulated data set, we computed the estimate of our target parameter $\Psi(P_0) \equiv EY_d$ for the following estimators: 1) Secant TMLE; 2) Iterative TMLE; 3) One-step TMLE; 4) Inverse Probability of Treatment Weighting (IPTW); 5) Efficient Influence Curve Estimating Equation Methodology (EE); 6) Maximum Likelihood Estimation using the G-computation formula. In the *Results* subsection we give bias, variance and relative MSE estimates. A brief description of each of the comparison estimators is given in chapter 1.

2.2.1 Data Generation

Please see Appendix B for a full description of the data-generation process for discrete (four-valued) $L(1)$.

2.2.2 Simulation Results

Estimates of bias, variance and relative mean squared error (Rel MSE) for all three parameters specified above are presented for the TMLE's and several

comparison estimators in tables 2.1 and 2.2. We defined estimated relative MSE for each estimator in (1.14).

Once again, the estimates of bias in all cases are not accurate to much less than 10^{-3}; we indicate estimates that appeared to be less than this with an asterisk.

Qm, gc denotes simulations where g (the treatment mechanism) was correctly specified, but $Q^0_{L(2)}$ was purposely misspecified. Qc, gc are simulations for which both Q and g are correctly specified. Note that the IPTW estimator is not affected by the form of Q^0 since this estimator does not estimate Q_0. Differences in IPTW performance between the sets of runs where Q is correctly specified and those where it is misspecified are thus the result of randomness in the simulations.

We generated 4000 independent simulations for each model specification/sample size combination (six sets of simulations in all).

2.3 Discussion

In chapter 1 we discussed the results as they pertain to the maximum likelihood-based TMLEs and the rest of the comparison estimators. We mention a few highlights here but focus on results from the secant-based TMLE.

In terms of the performance measures given in the tables, the differences between the three TMLEs implemented are insignificant. The one-step algorithm appears to hold a very slight bias advantage at the small sample size of 30 in estimating EY_1, but the relative MSE's are nearly the same. The overall performance of the secant TMLE could be improved slightly by intervening on simulation runs to ensure the algorithm converges. This makes the differences in bias that we report partly an artifact of the process of bias estimation by simulation, and not a true bias difference, assuming that in actual practice one can examine the empirical EIC for any given data set, which generally should be the case.

In almost every simulation the difference in estimates produced by the iterative and secant approaches was on the order of 10^{-4} or less, even at $n = 30$. The occasions in which the difference was significant were those in which one or the other algorithm failed to converge (in terms of yielding an estimate of Q that solved the empirical EIC) in the allotted number of steps.

The variance of the TMLE, EE and MLE estimators are already very close to the efficiency bound at $n = 100$ under Qc. In the chapter 1 simulations we found this to be the case for sample sizes of 250 and greater rather than at 100.

Qc, gc

n = 30

		Sec	Iter	1-step	IPTW	MLE	EE
EY_1	Bias	-0.028	-0.028	-0.023	-0.009	-0.041	-0.032
	Var	0.024	0.023	0.024	0.069	0.022	0.022
	Rel MSE	1.4	1.4	1.5	4.1	1.4	1.4
EY_2	Bias	-0.018	-0.018	-0.017	-0.004	-0.024	-0.019
	Var	0.027	0.027	0.028	0.062	0.027	0.027
	Rel MSE	1.4	1.4	1.4	3.0	1.3	1.3
EY_3	Bias	-0.017	-0.017	-0.017	-0.007	-0.024	-0.017
	Var	0.012	0.012	0.012	0.024	0.012	0.012
	Rel MSE	1.2	1.2	1.2	2.5	1.2	1.2

n = 100

		Sec	Iter	1-step	IPTW	MLE	EE
EY_1	Bias	-0.0019	-0.0019	-0.0020	*	-0.0017	-0.0021
	Var	0.0054	0.0054	0.0054	0.0212	0.0048	0.0054
	Rel MSE	1.1	1.1	1.1	4.2	1.0	1.1
EY_2	Bias	*	*	*	0.0016	*	*
	Var	0.0068	0.0068	0.0068	0.0197	0.0064	0.0067
	Rel MSE	1.1	1.1	1.1	3.2	1.0	1.1
EY_3	Bias	*	*	*	*	*	*
	Var	0.0032	0.0032	0.0032	0.0075	0.0031	0.0032
	Rel MSE	1.1	1.1	1.1	2.5	1.0	1.1

n = 200

		Sec	Iter	1-step	IPTW	MLE	EE
EY_1	Bias	*	*	*	*	*	*
	Var	0.0026	0.0026	0.0026	0.0103	0.0023	0.0026
	Rel MSE	1.0	1.0	1.0	4.1	0.9	1.0
EY_2	Bias	0.0015	0.0015	0.0016	0.0028	0.0010	0.0015
	Var	0.0032	0.0032	0.0032	0.0094	0.0029	0.0031
	Rel MSE	1.0	1.0	1.0	3.1	1.0	1.0
EY_3	Bias	*	*	*	0.0016	*	*
	Var	0.0014	0.0014	0.0014	0.0034	0.0014	0.0014
	Rel MSE	1.0	1.0	1.0	2.3	0.9	1.0

Table 2.1: *Qc, gc.* Estimator performance for various sample sizes with Q and g correctly specified, for each of three estimated parameters. The estimates for the iterative TMLE were from the 4th iteration. (*) indicates an estimated bias $< 10^{-3}$. (Based on 4000 simulations at each sample size.)

Qm, gc

n = 30

		Sec	Iter	1-step	IPTW	MLE	EE
EY_1	Bias	-0.0049	-0.0050	-0.0042	-0.0071	-0.2975	-0.0093
	Var	0.021	0.020	0.020	0.068	0.071	0.028
	Rel MSE	1.2	1.2	1.2	4.1	9.5	1.7
EY_2	Bias	-0.0010	*	*	0.0022	-0.1677	0.0090
	Var	0.024	0.024	0.024	0.064	0.075	0.029
	Rel MSE	1.2	1.1	1.2	3.1	5.0	1.4
EY_3	Bias	-0.0025	-0.0026	-0.0026	-0.0025	-0.2326	*
	Var	0.011	0.011	0.011	0.025	0.072	0.013
	Rel MSE	1.1	1.1	1.1	2.5	12.8	1.3

n = 100

		Sec	Iter	1-step	IPTW	MLE	EE
EY_1	Bias	-0.0040	-0.0040	-0.0044	0.0014	-0.3159	-0.0034
	Var	0.0056	0.0056	0.0056	0.0203	0.0326	0.0078
	Rel MSE	1.1	1.1	1.1	4.0	26.3	1.6
EY_2	Bias	0.0021	0.0021	0.0026	-0.0024	-0.1855	0.0026
	Var	0.0063	0.0063	0.0064	0.0187	0.0351	0.0080
	Rel MSE	1.0	1.0	1.0	3.0	11.3	1.3
EY_3	Bias	*	*	*	*	-0.251	*
	Var	0.0030	0.0030	0.0030	0.0072	0.0338	0.0036
	Rel MSE	1.0	1.0	1.0	2.4	32.6	1.2

n = 200

		Sec	Iter	1-step	IPTW	MLE	EE
EY_1	Bias	-0.0038	-0.0038	-0.0042	-0.0016	-0.3276	-0.0029
	Var	0.0028	0.0028	0.0028	0.0104	0.0187	0.0039
	Rel MSE	1.1	1.1	1.1	4.1	50.0	1.5
EY_2	Bias	0.0017	0.0017	0.0020	0.0012	-0.1962	0.0019
	Var	0.0033	0.0033	0.0034	0.0095	0.0200	0.0040
	Rel MSE	1.1	1.1	1.1	3.1	19.0	1.3
EY_3	Bias	-0.0010	-0.0010	-0.0011	*	-0.2620	*
	Var	0.0015	0.0015	0.0015	0.0034	0.0193	0.0017
	Rel MSE	1.0	1.0	1.0	2.3	59.4	1.1

Table 2.2: *Qm, gc.* Estimator performance for various sample sizes with Q misspecified and g correctly specified, for each of three estimated parameters. Estimates for the iterative TMLE were from the 4th iteration. (*) indicates an estimated bias $< 10^{-3}$. (Based on 4000 simulations at each sample size.)

The performance of the TMLE's at the small sample size of 30 is remarkable, particularly under model misspecification. Indeed, bias and variance of all three estimators are *better* when Q^0 is misspecified. The bias of the estimating equation estimator is also smaller under model misspecification. The advantage of the TMLEs' being substitution estimators also becomes apparent in these small sample results: at $n = 30$, many times the estimating equation and IPTW estimators gave estimates outside the range $[0, 1]$ even though the outcome is binary.

Misspecification of Q in all cases meant misspecifying $Q^0_{L(2)}$ but correctly specifying $Q^0_{L(1)}$. Thus under Qm, gc the MLE will be biased but the TMLE and EE estimators are double-robust and therefore still asymptotically unbiased under correct specification of g. Under the scenarios simulated here g is expected to be known and we therefore omitted simulations in which g is misspecified; the latter will of course result in bias of the IPTW estimator.

2.3.1 Convergence of the Secant Algorithm

In practice, we examined several plots of $f(\epsilon)$ vs ϵ to get a rough idea of its shape, and variability of shape, in order to select starting points for the secant algorithm in the simulations. Using these examples we selected fixed initial points for a given set of simulations, and never intervened on particular runs to ensure convergence of the algorithm. (The algorithm never failed to converge for $n \geq 100$.) The shape of most curves examined was made to order for the secant method, assuming well-chosen starting values (see below). In general practice, one would not need to specify starting points without first examining such a plot, which allows one simply to select starting points that will clearly lead to convergence. In effect, this just means selecting starting points that are "close enough" to the root. Unfortunately there is no generally agreed upon (or even proposed) notion of "close-enough" in the literature, but there are clear cases of it. For example, if the curve is roughly linear near the root, then starting points in the linear region will suffice.

There are also clear cases which can be problematic for finding the root in a reasonable number of iterations using the secant method. We have discovered two such general cases, both of which were observed only at the small sample size of 30. The first is when the curve has a point approaching zero slope between the two starting values (see figure 2.1). That is, assuming $f(\epsilon)$ is differentiable, then there is an $\epsilon' \in [\epsilon_0, \epsilon_1]$ such that

$$f'(\epsilon)\Big|_{\epsilon=\epsilon'} \approx 0.$$

(The empirical EIC is in fact differentiable for our parameters of interest.

More generally, if $P_n D^*(Q, g)$ were merely continuous and not differentiable at all points in the domain, then the situation above approximately corresponds to the existence of an ϵ' such that $\epsilon^* < \epsilon' < \epsilon_1$ or $\epsilon^* > \epsilon' > \epsilon_1$ implies $0 < |f(\epsilon')| \geq |f(\epsilon_1)|$, where ϵ^* is the root.) In fact the algorithm performs much worse if the position of zero slope is between ϵ_1 and ϵ^*, rather than between ϵ_0 and ϵ^*. Since this is true of the two starting points, (ϵ_0, ϵ_1), it is also true for the points $(\epsilon_k, \epsilon_{k+1})$ corresponding to the k^{th} iteration of the algorithm.

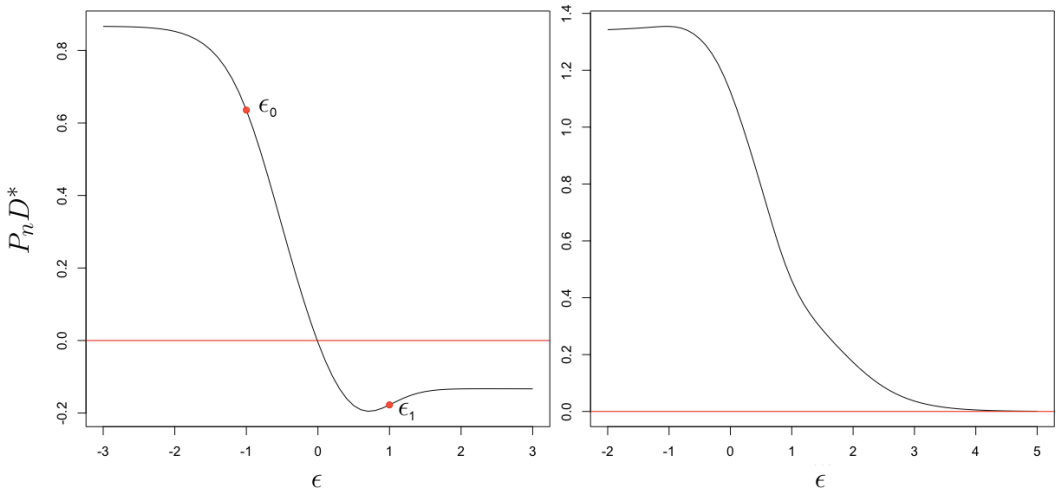

Figure 2.1: Two examples of $P_n D^*(\epsilon)$ for which the secant method failed to converge at $n = 30$. **Left:** Point of zero slope in starting interval. No convergence with the two indicated starting values (-1,1) in 10 steps or less. Starting values $(\epsilon_0, \epsilon_1) = (0.25, 0.5)$ did yield converence. **Right:** Curve approaches 0 slowly near the root. The true ϵ_s in this case was ≈ 20.81. Starting values (-0.5, 0.5) failed to converge but alternate starting points did yield convergence.

The second difficulty arises when $f(\epsilon)$ approaches zero slowly near the root (see figure 2.1). In this case the secant method is known to have trouble converging in a reasonable number of iterations even if the starting values yield a value of $f(\epsilon)$ that is relatively close to zero. Several of our simulations at $n = 30$ confirm this. Interestingly, these tend also to be cases in which all the TMLE's give a parameter estimate of either 1 or $1 - \delta$ with $\delta < 0.05$. In these cases, the TMLE is trying to force the estimate to 1. Regardless of the reason for this, the situation is reflected in the empirical EIC, which reveals that the solution ϵ_s, is relatively far from 0. Since ϵ_s is the coefficient in front of the clever covariate term, a large (absolute) value of ϵ_s (assuming a non-negligible clever covariate) will result in a large term in the exponential expression in the denominator of $Q_n(\epsilon)$, and drive $\Psi(Q_n(\epsilon))$ toward 0 or 1. Nevertheless, even in these cases the secant-based TMLE tends to agree with the maximum likelihood-based TMLE's—they all produce estimates very close to 1. These are cases in which the TMLE methods are breaking down due to sparsity. (They are not cases of positivity violation, since g_0 is given and bounded well away from 0.)

Despite these potentially problematic types of curves, the fact that the secant method is not guaranteed to converge in general appears to be no drawback at all in our situation. One can always examine $f(\epsilon)$ in the neighborhood of the root and pick initial estimates in an informed way—i.e., close enough to the root to avoid the potential problems described above. We were able to do this whenever the initial starting values did not result in convergence to a solution in ten iterations of the algorithm or less. It may be that there are cases in which the empirical EIC behaves so poorly in the neighborhood of the root that this technique fails when there is in fact a solution, but we observed no such cases.

There are also so-called "safeguarded" algorithms which force each iteration to bracket the solution by ensuring that the two current estimates are of opposite sign. The method of false position is one such algorithm (Faires and Burden, 2003). Such methods can be used to guard against divergence of the method, and can be used in place of the secant method if for some reason an a priori guarantee of convergence is required.

2.3.2 Comparison of One Step, Iterative and Numerical Solution TMLE Algorithms

All targeted minimum loss-based estimators—including the "numerical methods" TMLE—are double-robust, and are efficient under correct model specification.

The advantage of the the numerical methods approach (secant, bisection, Newton, etc.) is that it is the easiest overall to implement, given $K \geq 2$ (where K is the number of time-points at which data is measured). Next in terms of implementation complexity is the one-step algorithm, and finally the iterative approach.

Also noteworthy is that the one-step TMLE requires estimation of two ϵ's in the binary $L(1)$ case and four ϵ's when $L(1)$ has four levels (three for L(1) and one for L(2)). For the general data structure $(L(0), A(0), ...L(J), A(J), L(J+1))$ where intermediate factor $L(j)$ has t_j levels, the number of ϵ's the one-step estimator must fit is $\sum_{j=1}^{J+1}(t_j - 1)$. In contrast, the iterative and numerical solution TMLE's perform a fitting of a single ϵ. It would therefore not be surprising to see at least a small efficiency advantage for the iterative and numerical methods as K and/or t_j increase, though we have not observed any such advantage in the present simulations.

The three methods also differ slightly in terms of computational resources required. For the data simulated here, at $n = 2000$, the order in terms of computational speed was 1) one-step, 2) iterative and 3) secant. However, this result was based on running four iterations of the iterative procedure and

imposing a tolerance $|P_n D^*(Q_n(\epsilon))| \leq 10^{-6}$ on the secant algorithm, both of which criteria are overkill. Since typically for the k^{th} iteration of the secant method, $|\psi_k - \psi_{k+1}| \approx |P_n D^*(\epsilon_k)|$, a reasonable tolerance is, say, $|P_n D^*(\epsilon_k)| \leq var(\psi_k)/10$. Such a tolerance will make the speed of the secant procedure comparable to the iterative procedure.

It is possible that under some conditions the empirical EIC has multiple solutions, or no solution, though we observed no such cases. If multiple solutions, one could select the ϵ_s that yielded the highest likelihood. If the EIC has no solution for the single ϵ approach then the one-step procedure would be favored.

Chapter 3

A Comparison of TMLEs in the Longitudinal Setting

3.1 Introduction

We have already seen that targeted minimum loss-based estimation is a highly flexible substitution estimation methodology that is locally efficient and double-robust. Indeed, estimating parameters associated with longitudinal data allows for many choices of TMLE's, and the investigator faces many more choices even after choosing a particular TMLE. In this chapter we investigate a comparison between two general categories of TMLE, and within each, a set of TMLEs indexed by different initial estimators of the relevant conditional densities and/or conditional means expressed in the G-computation formula that represents the parameter to be estimated.

We focus on the same general data structure that is defined in chapter 1, though this time we choose an outcome that is continuous on the interval [0,1].

Again we take as the parameter(s) to be estimated a set of treatment-specific means, as defined in chapter 1. Recall that in the sequentially randomized controlled trial setting, the treatment mechanism, $g \equiv P(A = a \mid Pa(A))$, is known and therefore double-robust estimators, including all TMLE's, and inverse probability of censoring weight (IPCW) estimators are guaranteed to be consistent estimators of the relevant parameter. The double-robustness property of the TMLE allows for a number of options in estimating the non-g part of the likelihood, which we call Q:

$$p(O) = Q(O)g(O)$$

where, to reiterate,

$$Q \equiv \prod_j Q_j \equiv \prod_j P(L(j) \mid \bar{L}(j-1), \bar{A}(j-1) \text{ and}$$
$$g \equiv \prod_j g_j \equiv \prod_j P(A(j) \mid \bar{L}(j), \bar{A}(j-1)$$

Since consistency is guaranteed under correct specification of g in a double-robust estimator, efficiency is a driving force in choosing an approach to estimation of Q, assuming finite sample bias is more or less equal among the set of competing estimators.

The focus of this chapter is two-fold. First we interest ourselves in the relative performances of two classes of TMLE's. Each general class attempts to estimate the factors of a so-called g-computation formula, which is a representation of the counterfactual parameter of interest in terms of the observed data. The first class, which we shall call a *density based-TMLE (db-TMLE)* corresponds with a G-computation formula that consists of a series of conditional densities, and is the subject of the first two chapters of this disseration; the second corresponds with a formula that only includes a set of nested conditional expectations (*nce-TMLE*). van der Laan and Gruber (2012) describe this latter TMLE, which represents a fundamentally different approach to producing a TMLE than the former. The latter has been theorized to have advantages over estimators that must estimate a larger portion of the likelihood, possibly including the db-TMLEs.

The second aim is to investigate the performance of each class under different initial estimators of the relevant factors of the respective G-computation formulas. (The TMLE method proceeds by obtaining an initial estimator of the relevant factor and then updating this initial estimate by fluctuating it along a least favorable parametric submodel. See van der Laan and Rubin, 2006.) In the first two chapters we observed the behavior of the TMLEs under correct and incorrect model specification. Correct model specfication is not of particular interest to the investigator who seeks to estimate the effect of some real world treatment or exposure, since he or she will not know the correct model. Incorrect specification is likely, but, as we remarked earlier, we're interested in how the estimators perform under an initial model specification that is likely to be employed. We therefore would like to observe the behaviors using data-adaptive estimation of the initial estimators of the parameters of the G-computation formula.

After some initial exploration, we believe the nce-TMLE shows more promise in its ability to minimize MSE in estimating the parameter of interest, and consequently we explore an additional approach to initial estimator selection for this TMLE. It remains for future work to perform a similar comparison for the db-TMLE.

Chapters 1 and 2 together describe three db-TMLEs for the longitudinal setting, all of which perform comparably. We here select the so called one-step method to represent the db-TMLEs in the current simulation study.

3.2 Parameter of Interest and G-computation Formulas

We are interested in the same treatment-specific mean as has been described in earlier chapters.

3.3 Two Classes of TMLEs

3.3.1 Density-Based TMLE (db-TMLE)

The db-TMLEs we've described and implemented thus far attempt to estimate the parameters of the G-computation formula given in (1.4).

Note that the second and third multiplicative terms in the summand are densities. While $P(L(0) = l(0))$ can be estimated non-parametrically, estimating the conditional density of $L(1)$ requires the estimation of the conditional probability of $L(1) = l(1)$ for each possible value of $l(1)$. In the present case this amounts to the estimation of four conditional probabilities, which can be achieved as was mentioned in chapter 1:

Code each of the categories for $L(1) \in \{0, 1, 2, 3\}$ as a binary indicator variable, $L(1, m),\ m = 0, 1, 2, 3$, and relate them to the categorical variable $L(1)$ as shown in (1.8). As we mentioned, in this way the conditional density of each binary factor of $L(1)$, $Q_{L(1,m)}$, can be estimated using logistic regression. (Further details of the implementation of this approach were given in chapter 1.)

We mentioned at the beginning of this chapter that one can index any TMLE by initial estimators of the parameters of the G-computation formula. In the case of the db-TMLE this means one has a choice of initial estimators of the relevant densities. Our preferred approach for initial estimation is the highly data-adaptive Superlearner algorithm (van der Laan et al., 2007b). Naturally one will not know the true $Q_0(P)$ that generated the data, but we include the correct model as an initial estimator here for comparison against the more realistic case of using data-adaptive methods to construct an initial estimator.

3.3.2 Nested Conditional Expectation TMLE (nce-TMLE)

By the so-called iterative conditional expectation or "tower" rule, EY^d can also be written as a series of nested conditional expectations. The inner-most such conditional expectation is

$$\begin{aligned} & E\left(Y \mid \bar{A} = d(\bar{L}), \bar{L}(J)\right) \\ =& E\left(Y \mid Pa^d(Y)\right). \end{aligned} \tag{3.1}$$

Here we introduced the notation $Pa^d(X(j))$ to indicate the set $Pa(X(j))$ with each intervention node set to the value dictated by the specified intervention. We use the notation $\bar{Q}^d_{J+1}$ to signify this innermost conditional expectation. Similarly, define

$$\bar{Q}^d_J \equiv E\left(\bar{Q}^d_{J+1} \mid Pa^d(L(J))\right)$$

and so on for $\bar{Q}^d_j$, $j = J-1, ...0$. The parameter we want to estimate can be written in terms of these nested conditional expectations where for each j, the relevant conditional expectation is taken conditional on $\bar{A}(j-1) = \bar{d}_{j-1}, \bar{L}(j-1)$. We eventually reach the last conditional expectation in the series, and finally a marginal expectation. With the above notation, we're able to express the parameter starting at any point in the series. For example,

$$\begin{aligned} \Psi_d(P) &= \Psi(\bar{Q}^d) \\ &= \bar{Q}^d_0 \\ &\equiv E[E\left(\bar{Q}^d_1 \mid Pa^d(L(1))\right)] \\ &\equiv E[E\left(E(\bar{Q}^d_2 \mid Pa^d(L(2)) \mid Pa^d(L(1))\right)], \end{aligned} \tag{3.2}$$

and so on, where $\bar{Q}^d \equiv \left(\bar{Q}^d_{J+1}, \bar{Q}^d_J, ..., \bar{Q}^d_0\right)$.

The nce-TMLE algorithm begins by obtaining an initial estimator of $\bar{Q}^d_j$ for each j; we then update this estimator with a fluctuation function that is tailored specifically to remove bias in estimating the parameter of interest. The TMLEs do not dictate the use of any particular estimation method for $\bar{Q}^d_j$, though there are clear gains in finite sample performance if $\bar{Q}^d_{j,n}$ is close to $\bar{Q}^d_j$.

As with the db-TMLEs, upon obtaining the initial estimates of $\bar{Q}^d_j$, these esti-

mates are fluctuated such that the fluctuated updates of the initial estimates are guaranteed to yield a consistent estimate of ψ_d if either the initial estimates, $\bar{Q}^d_{j,n}$, are consistent estimators of $\bar{Q}^d_j$ or our estimates $g_{j,n}$ of g_j for each j are consistent, i.e., like all TMLEs, this TMLE is double-robust. And, as usual, if both estimates are consistent, then the estimator is asymptotically efficient.

Suppose for example that $\bar{Q}^d_{J+1,n}$ is an initial estimate of $\bar{Q}^d_{J+1}$ obtained via some preferred approach (e.g., logistic regression). The updated estimate, which, approximately following van der Laan and Gruber (2012), we designate $\bar{Q}^{d,*}_{J+1,n}$ is obtained by using the initial estimator as an offset in a univariate logistic regression of Y on the covariate $c(J+1)$, where

$$c(J+1) = \frac{I(\bar{A}(J) = \bar{d}_J)}{\prod_{j=0}^{J} g_j} \tag{3.3}$$

The relevant covariates for $1 < j < J+1$ are analogous. In these simulations, the correct g was used in computation of these covariates in implementing the estimator.

The estimation of each successive conditional expectation starting with $\bar{Q}^d_{J+1,n}$ is updated in this manner yielding the corresponding $\bar{Q}^{d,*}_{j,n}$ for $j = J+1, J, ..., 1$. The parameter estimate is then computed as $\psi_{d,n} \equiv \Psi(\bar{Q}^{d,*}_n) = \bar{Q}^{d,*}_{0,n}$

We include comparisons of the nce-TMLE using both Superlearner and the "correct" model as initial estimators of $\bar{Q}^d$, as well as the initial estimators described below. (See section 3.4 for an explanation of the quotation marks.)

Additional Targeting of the nce-TMLE

Since in an SRCT the treatment mechanism, g, is known, double robust estimators present the opportunity to make choices for estimation of $\bar{Q}^d$ that might further improve efficiency. We've mentioned TMLE's indexed by the Superlearner-based initial estimator and the parametric model-based initial estimator.

Note that the Superlearner uses average cross validated risk, with squared error as the loss function, as a criterion in building its overall prediction model. However, one can use as loss function the variance of the efficient influence curve itself—an excellent choice since it targets ψ_0 directly (see van der Laan, 2010a) because it corresponds with the asymptotic variance of the TMLE:

$$\mathcal{L}(\bar{Q}) = D^*(\bar{Q}^d, g_0)^2 \tag{3.4}$$

This loss function is valid since it satisfies $\bar{Q}_0^d = argmin_{\bar{Q}^d} E_0 \mathcal{L}(\bar{Q}^d)(O)$ among all $\bar{Q}^d$ for the parameter $\Psi(\bar{Q}_0^d) = \psi_d$. One could thus select among initial estimators by picking the one that minimizes the risk corresponding to this loss function. Indeed, one could employ superlearning using the above as loss function, which is a good direction for future research. In our simulations we in fact implemented two distinct nce-TMLEs that use either 1) the entire efficient influence curve (EIC) or 2) separate components of it as loss functions, and select the TMLE with the lowest empirical variance of the estimated EIC.

The representation of the EIC for parameter EY^d in our two time-point longitudinal setting, in terms of nested conditional expectations, is the sum of three random variables (van der Laan and Gruber, 2012). Defining first

$$c(j) = \frac{I(\bar{A}(j-1) = \bar{d}_{j-1})}{\prod_{s=0}^{j-1} g_s}, \tag{3.5}$$

$$\tag{3.6}$$

for $j = 1, 2$, we have that the EIC can be written

$$D^*_{\psi_d}(\bar{Q}^d, g) = \sum_{j=0}^{2} D^*_{\psi_{d,j}}$$

with

$$\begin{aligned} D^*_{\psi_{d,2}} &= c(2)(Y - \bar{Q}_2^d) \\ D^*_{\psi_{d,1}} &= c(1)(\bar{Q}_2^d - \bar{Q}_1^d) \\ D^*_{\psi_{d,0}} &= \bar{Q}_1^d - \Psi(\bar{Q}^d). \end{aligned}$$

1. Selection Based on Minimizing the Entire EIC

In the following we suppress the superscript d and note that whenever $\bar{Q}$ appears we mean $\bar{Q}^d$. Given a collection of candidate estimators $\bar{Q}_k$ of $\bar{Q}$, one way of using the EIC loss function to select k from K possible estimators is with the cross-validation selector k_n

$$k_n = k(P_n) = \underset{k}{argmin} E_{B_n} P^1_{n,B_n} D^*(\bar{Q}^*_k(P^0_{n,B_n}), g_0)^2. \tag{3.7}$$

Here $B_n \in \{0,1\}^n$ is a random vector of binary variables determining the split into training samples ($\{i : B_n(i) = 0\}$) and validation samples ($\{i : B_n(i) = 1\}$). P^0_{n,B_n} and P^1_{n,B_n} are the empirical distributions of the training and validation samples, respectively, for a given split B_n. We also used the notation $P_n f \equiv 1/n \sum_{i=1}^n f(O_i)$. In words, implementing this selector corresponds to splitting the data into, say, V pairs of training and validation samples. Let R be the number of learners considered for estimation of $\bar{Q}_2$. For each split B_n, one trains the candidate $\bar{Q}_{2,r}$ on P^0_{n,B_n} and uses the resulting prediction model to obtain estimates of $\bar{Q}_2(O_i)$ for $\{i : B_n(i) = 1\}$. This is done for each split $v \in 1, 2, ..., V$ such that one eventually has a cross-validated estimate of $\bar{Q}_2(O)$ for each observation i in the data set. TMLE updating of these initial estimators is done as usual to produce the r-specific cross-validated TMLE $\bar{Q}^*_{2,r_{cv}}$.

One now proceeds likewise for $j = 1$. However, since now there are R possible distinct estimators of $\bar{Q}_2$, each of them must be used in turn for the cross-validated fitting of each of the (say S) estimators of $\bar{Q}_1$ thus giving a total of $K = S * R$ such estimators. Finally, each fluctuation-updated $\bar{Q}^*_{k_{cv}}$ yields a corresponding cross-validated TMLE $\psi_{k_{cv}} = \Psi(\bar{Q}^*_{k_{cv}})$, which is needed in the computation of the associated estimate of the cross-validated influence curve.

This procedure is performed for each candidate estimator $\bar{Q}_k$. Once the cross-validation selector (3.7) has been applied, the selected algorithm $\bar{Q}_{k_n}$ is fitted on the full data set to yield the cross validation selector-based TMLE $\Psi(\bar{Q}^*_{k_n})$.

As we hint at above, $D^*_{\psi_d}(\bar{Q}^d, g) = D^*_{\psi_d}(\bar{Q}^d_1, \bar{Q}^d_2, g)$, i.e., $\bar{Q}^d$ is composed of two parts; thus each k-specific estimator $\bar{Q}_k$ consists of two algorithms, one for each $j = 1, 2$. Thus if one has S candidate estimators of $\bar{Q}^d_1$ and R estimators of $\bar{Q}^d_2$ then one is selecting among $K = S * R$ estimators of $\bar{Q}^d$. Since one must compute $\psi_{k_{cv}}$ for each k, K cross-validated TMLEs must be evaluated before the final estimate $\Psi(\bar{Q}^*_{k_n})$ is obtained. This fact has computational ramifications, especially as one considers data with more than two treatment or exposure times. Even for $J = 2$, in our simulations $S \times R = 14 \times 18 = 252$ algorithms and we consider this S and R relatively small compared to what is possible with this technique.

2. Selection Based on Minimizing the Components of the EIC Separately

Another way of incorporating the EIC into a targeted loss function is to apply a cross-validation selector based on minimizing the variance of each component

of the EIC separately. Recalling that the EIC for our parameter of interest has $J+1$ orthogonal components, the corresponding selector is

$$k_n = k(P_n) = \underset{k}{argmin} \sum_{j=0}^{J+1} E_{B_n} P^1_{n,B_n} D_j(g_{j,0}, \bar{Q}^*_{j,k_j}(P^0_{n,B_n}))^2 \tag{3.8}$$

where $g_{j,0}$ is the correctly specified treatment mechanism for the j^{th} time point, and k_j now denotes the k^{th} estimator out of the K_j considered for estimating $\bar{Q}_j$. In practice, the sum is only taken over $j = 1, 2$ since the final component of the EIC consists in the quantity $\bar{Q}_1 - \Psi(\bar{Q})$, and the estimator of $\bar{Q}_1$ has already been selected based on minimizing the variance of D^*_1. Since this procedure does not necessarily select the TMLE of ψ with the lowest estimated overall variance, one would expect it to not perform as well as the first cross-validated selector described above.

Since one is minimizing the sum of two positive entities, the minimum for each component of the EIC can be found separately. First one chooses the r_n such that

$$r_n = \underset{r}{argmin} E_{B_n} P^1_{n,B_n} D_2(g_{2,0}, \bar{Q}^*_{2,r}(P^0_{n,B_n}))^2. \tag{3.9}$$

(In keeping with our notation above, we assume here $K_2 = R$ candidate estimators for $\bar{Q}_2$ and $K_1 = S$ candidates for $\bar{Q}_1$.) One then fits $\bar{Q}_{r_n}$ on the full data (or the subset for which $\bar{A}(2) = \bar{d}(2)$, depending on how one chooses to estimate $\bar{Q}^d_2$), performs the TMLE update, and uses $\bar{Q}^*_{r_n}$ as "outcome" for the next stage of the NCE algorithm, viz., selection of $\bar{Q}_{1,s_n}$. The cross-validation selector for the next initial estimator, s_n is found in the same manner as depicted in (3.9), but with D_1 in place of D_2, and with the appropriate subscripts on $\bar{Q}^*$ and g. To compute the final estimation of ψ_d one now starts the algorithm from the top, but uses the overall k-specific cross-validation selector $k_n(r_n, s_n)$ as initial estimator. In this case, assuming one chooses from R candidates for $\bar{Q}_2$ and S candidates for $\bar{Q}_1$, there are still a total of $R * S$ possible estimators of $\bar{Q}$, but since the cross-validation selectors are implemented sequentially, one only has to perform $R + S$ cross-validated fittings. This procedure is thus computationally much quicker than the first selector described.

3.3.3 Other Comparison Estimators

Data-adaptive G-computation Estimation

This a straightforward data-adaptive estimator of the parameters of the G-computation formula given in (1.4). We use the Superlearner here as well. These initial estimates of the densities are also the final estimates, i.e., there is no fluctuation updating.

Using data-adaptive estimation is a good step away from parametric modeling, but a major drawback (aside from being biased if one does not know the correct parametric model) is that this estimation method does not produce an asymptotically linear estimator and thus one cannot use standard methods to compute confidence intervals (e.g., by computing the variance of the estimator's influence curve). No theory yet exists that establishes that the estimator converges as n increases without bound, which means the validity of even the bootstrap for obtaining inference is not known.

IPTW Estimators

The basic IPTW estimator of the parameter EY^d was described in chapter 1:

$$\psi_n^{IPTW} = \frac{1}{n}\sum_{i=1}^{n} Y_i \frac{I(\bar{A}_i = d(\bar{L})}{g[\bar{A}_i = d(\bar{L}) \mid X_i]}.$$

The so-called *stabilized* IPTW, which typically performs much better than the simple IPTW estimator (and certainly does in these simulations), is obtained by dividing the above estimator by the average of the weights:

$$\psi_n^{stb-IPTW} = \frac{\sum_{i=1}^{n} Y_i \frac{I(\bar{A}_i=d(\bar{L})}{g[\bar{A}_i=d(\bar{L})|X_i]}}{\sum_{i=1}^{n} \frac{I(\bar{A}_i=d(\bar{L})}{g[\bar{A}_i=d(\bar{L})|X_i]}} = \frac{\psi_n^{IPTW}}{\frac{1}{n}\sum_{i=1}^{n} \frac{I(\bar{A}_i=d(\bar{L})}{g[\bar{A}_i=d(\bar{L})|X_i]}}.$$

3.4 Results and Discussion

The data-generating functions for the simulations are given in Appendix C. Note that, in contrast to the simulations for chapters 1 & 2, this time $Y \in [0, 1]$. This allows one more control of $var(Y)$ in designing data generation functions, and thus the degree of predictability of the outcome from the covariates. The EIC tells us that the more predictable the outcome is (as a function of $Pa(Y)$) the smaller the variance of the EIC is. This has implications for the relative efficiency of an estimator whose influence curve is the EIC vs. an estimator's whose influence curve is not.

Table 3.1 lists the various estimators and their performances in terms of bias, relative efficiency and relative mean squared error (MSE), the latter two of which were defined in chapter 1. We are estimating the same three parameters as previously, i.e., EY_d for three treatment rules $d = 1, 2, 3$.

3.4.1 TMLEs Using Superlearner

The first comparison of interest for us is between the density-based and nested-conditional-expectation TMLEs. In estimating all parameters and at all sample sizes, the nce-TMLE using the full dataset for fitting the initial $\bar{Q}$'s outperforms the db-TMLE in terms of MSE when using Superlearner for estimating initial $\bar{Q}$. The gains are more in efficiency than in bias. Of significant interest as well is that the nce-TMLE that uses the entire dataset for fitting of initial $\bar{Q}$ performs substantially better than the version that fits $\bar{Q}$ on the subset of the data which follow the relevant intervention rule.

The latter nce-TMLE sometimes outperforms the db-TMLE, and sometimes not. At $n = 100$, db prevails while at $n = 500$ the nce version does a little better on two of the three parameters. Among TMLEs then, when using Superlearner for initial estimation of the relevant parameter (conditional density or conditional expectation), the nce version using the full dataset for fitting Superlearner is the clear winner in estimating the specified parameters and for this particular data-generating distribution. Indeed, in most cases this estimator achieves a variance below the semi-parametric efficiency bound.

It's also interesting to note that for all parameters and at all sample sizes, the SL-based nce-TMLE performs nearly the same as when using the "correct" model specification for $\bar{Q}$. "Correct" is in quotes because we do not know the correct form of $E(\bar{Q}_Y^d \mid A(0) = d_0, L(0))$. Instead we approximate it by $E(L(1) \mid A(0) = d_0, L(0))$, since that function is used to generate the data. This is especially noteworthy given the small number of learners in the SL library: 15 for $\bar{Q}_2$ and 11 for $\bar{Q}_1$, many of which were parametric glms. The lack of a completely correct specification of $\bar{Q}_1^d$ may partially explain the close match between SL and "correct" model nce-TMLE's. On the other hand, there is a noticeable difference between correct model and Superlearner based db-TMLEs.

3.4.2 TMLEs Using Variance of the Efficient Influence Curve as Loss Function

As mentioned in our descriptions of the estimators, another nce-TMLE of particular interest to us is that which uses the variance of the EIC as loss function in selecting the initial $\bar{Q}$. There are four nce-TMLEs we examined.

n = 100

		db-TMLE		nce-TMLE (full)				nce-TMLE (sub)				Comparison		
		SL	*Qc*	SL	Qc	sep D_j^*	Full D^*	SL	Qc	sep D_j^*	Full D^*	IPTW	IPTW (*stb*)	SL G-comp
EY_1	bias	0.0015	*	-0.0027	*	-0.0174	-0.0030	-0.0023	*	-0.0031	-0.0012	-0.0017	*	-0.0161
	Rel Eff	0.97	0.89	0.75	0.75	0.87	0.82	1.02	0.96	1.13	0.97	55.01	0.95	1.47
	Rel MSE	0.98	0.89	0.77	0.75	1.85	0.85	1.04	0.96	1.16	0.97	55.02	0.95	2.31
EY_2	bias	-0.0033	*	-0.0019	*	0.0181	0.0118	*	*	0.0094	0.0062	0.0031	0.0016	-0.0069
	Rel Eff	1.10	0.97	0.89	0.89	0.51	0.81	1.12	1.02	1.23	1.17	14.35	1.15	0.88
	Rel MSE	1.11	0.97	0.89	0.89	0.85	0.95	1.12	1.02	1.32	1.21	14.36	1.16	0.93
EY_3	bias	*	*	*	*	-0.0016	*	0.0014	*	0.0012	0.0031	*	*	-0.0114
	Rel Eff	0.91	0.86	0.73	0.72	0.82	0.76	1.06	1.08	1.10	1.14	11.85	1.10	1.06
	Rel MSE	0.91	0.86	0.73	0.72	0.83	0.76	1.07	1.08	1.10	1.17	11.85	1.10	1.43

Table 3.1: Performance at $n = 100$ of 10 TMLEs and three comparison estimators. *nce-TMLE (full)* indicates nested conditional expectation TMLE where fitting of each initial estimator of $\bar{Q}$ was done on the entire dataset. *sub* means the fitting was done among observations following the intervention up to the relevant time point. *sep* D_j^* indicates the TMLE which uses the variance of the separate components of the EIC as loss function to select among candidate estimators, and *Full* D^* uses the full EIC at once. Relative efficiency is defined as $RE \equiv \frac{var(\psi_n)}{var(D^*(\bar{Q}_0, g_0))/n}$ and relative MSE as Rel MSE $\equiv \frac{(E(\psi_n)-\psi_0)^2+var(\psi_n)}{var(D^*(\bar{Q}_0, g_0)/n}$. Asterisks indicate bias < 0.001.

n = 200

		db-TMLE		nce-TMLE (full)				nce-TMLE (sub)				Comparison		
		SL	Qc	SL	Qc	sep D_j^*	Full D^*	SL	Qc	sep D_j^*	Full D^*	IPTW	IPTW (stb)	SL G-comp
EY_1	bias	0.002	*	-0.0019	*	-0.0175	-0.0018	-0.0019	*	-0.0027	-0.0012	-0.0027	*	-0.0116
	Rel Eff	1.01	0.92	0.77	0.76	1.24	0.77	1.01	0.99	1.06	0.99	61.06	0.99	1.38
	Rel MSE	1.04	0.92	0.79	0.76	3.22	0.79	1.03	0.99	1.11	1.00	61.11	0.99	2.24
EY_2	bias	-0.003	*	-0.0024	*	0.0180	0.0088	*	*	0.0201	0.0045	-0.0035	0.0014	-0.0031
	Rel Eff	1.31	1.11	1.00	1.00	0.59	0.93	1.16	1.13	1.79	1.21	13.74	1.33	0.94
	Rel MSE	1.33	1.11	1.01	1.00	1.27	1.09	1.16	1.13	2.63	1.25	13.76	1.33	0.96
EY_3	bias	*	*	*	*	-0.0014	*	0.0012	*	*	0.0023	*	*	-0.0072
	Rel Eff	1.08	1.04	0.84	0.84	0.92	0.83	1.29	1.26	1.30	1.31	13.00	1.30	1.04
	Rel MSE	1.08	1.04	0.84	0.84	0.93	0.83	1.30	1.26	1.30	1.34	13.00	1.30	1.34

Table 3.2: Performance at $n = 200$. Asterisks indicate bias $<$ 10e-3.

n = 500

		db-TMLE		nce-TMLE (full)				nce-TMLE (sub)				Comparison		
		SL	Qc	SL	Qc	sep D_j^*	Full D^*	SL	Qc	sep D_j^*	Full D^*	IPTW	IPTW (*stb*)	SL G-comp
EY_1	bias	0.0032	*	*	*	-0.02	*	-0.0012	*	-0.0022	*	0.0032	*	-0.0077
	Rel Eff	1.05	0.88	0.74	0.74	1.50	0.76	0.91	0.91	0.91	0.92	59.27	0.94	1.24
	Rel MSE	1.22	0.88	0.75	0.74	7.98	0.77	0.94	0.92	0.99	0.93	59.43	0.94	2.19
EY_2	bias	-0.0037	*	-0.0028	*	0.0162	0.0063	*	*	0.027	0.0017	-0.0037	*	*
	Rel Eff	1.27	1.07	0.97	0.98	0.58	0.98	1.09	1.09	2.13	1.15	12.7	1.29	0.84
	Rel MSE	1.34	1.08	1.01	0.98	1.97	1.19	1.10	1.09	5.96	1.16	12.78	1.29	0.84
EY_3	bias	*	*	*	*	-0.002	*	*	*	-0.0029	0.0018	*	*	-0.0044
	Rel Eff	1.00	0.97	0.78	0.78	0.84	0.79	1.18	1.15	1.38	1.24	11.8	1.19	0.82
	Rel MSE	1.00	0.97	0.79	0.79	0.90	0.79	1.18	1.15	1.50	1.28	11.81	1.20	1.09

Table 3.3: Performance at $n = 500$. Asterisks indicate bias $<$ 10e-3.

For each of the two ways of fitting the initial $\bar{Q}$ (using the full dataset, or the subset which followed the intervention), there were the two ways of utilizing the EIC as loss function: selecting each $\bar{Q}_j$ separately, based on minimizing $var(D_j^*)$, or selecting the overall $\bar{Q}$ that minimizes $var(D^*)$.

Overall this MVE (Minimum Variance of the EIC) nce-TMLE did not outperform the SL-based nce-TMLE in terms of MSE, though the estimator that used the full EIC came close. The comparison between SL-based and MVE TMLEs in these simulations is particularly apt since they both incorporated approximately the same number and type of learners. Indeed, one can consider the method used to select the initial estimator a Superlearner-based approach with loss function as specified in (3.4), though only a single k-specific learner is selected, rather than a convex combination of learners, as is the case with the Superlearner algorithm utilized in the SL-based TMLEs. This Superlearner-based initial estimator thus is at a slight disadvantage compared to what it might achieve if the learners could be combined as in the Superlearner package implemented in R.

There were notable cases in which the MVE TMLE variance was somewhat lower, but usually at the expense of some bias, which resulted in a higher MSE. For example, at all sample sizes and using the full data for fitting initial $\bar{Q}$, the MVE nce-TMLE that utilized minimization of the variance of the D_j^* separately had a remarkably low variance in estimating EY_2. Yet the bias of this estimator, though small, does not appear to decrease with sample size.

Though this is indeed a TMLE, and all TMLEs are expected to be asymptotically unbiased under correct specification of g, this property can be defeated in a number of ways. One way in which it may have been defeated here is if the cross-validation selector selected a model that predicted the same outcome for all observations that followed d. Suppose for example that the selected model $\bar{Q}_{1,k_n}$ is only a function of $A(0)$. Then for EY_1, EY_2 the values $\bar{Q}_{1,k_n}(O_i)$ generated by this prediction model will be equal for all observations that followed d_0 since for EY_1, $d_0(L(0)) = 1$ and for EY_1, $d_0(L(0)) = 0$. Furthermore, since $A(0)$ was randomized with $P(A(0) = 1) = 0.5$, all observations which followed d for $EY_{d=1}, EY_{d=2}$ will also have the same computed value of the clever covariate $c(1)$ (as defined in (3.5))using correct specification of g. Therefore the TMLE updating step associated with this initial $\bar{Q}_{1,k}$ will do nothing (same offset and same clever covariate for all observations in the set where $A(0) = d_0(L(0))$), and one has produced a TMLE that has a misspecified $\bar{Q}_1$ and no effective fluctuation updating, i.e., an estimator that will not generally remove the bias associated with model misspecification. This particular scenario did indeed obtain in some of the simulations.

Such a scenario is less likely using the Superlearner with standard squared error loss function since, as noted above, the SL will typically yield a convex combination of several learners for its prediction model.

As with the SL and correct model-based nce-TMLEs, the MVE nce-TMLEs that used the full data for fitting the initial $\bar{Q}$ outperformed those that used the subset which followed the intervention for initial fitting. Further, with one exception, the MVE nce-TMLE that used the full EIC as loss function outperformed the version minimizing the variance of the D_j^* separately. The exception was in estimating parameter EY_2 at $n = 100$. Here the extremely low variance of the latter separate D_j^* estimator dominates the MSE term and the MSE beats all estimators including both TMLEs that used the correct model as initial estimator. As n increases however, the bias of this estimator starts to dominate in the MSE and it loses out to the full EIC loss function version.

The MVE nce-TMLEs would have been improved had the Superlearner itself been one of the K candidate algorithms in the candidate library, but run time considerations prevented implementation of this in the present study.

We conclude from these comparisons that, for this particular set of data-generation mechanisms, the SL-based nce-TMLE is the overall winner. The MSE of the estimator was consistently lower than all other estimators, with the occasional exception of the MVE nce-TMLE. And even in the latter cases the differences in MSE are insignificant.

3.4.3 Comparison Estimators

The non-stabilized version of the IPTW performs terribly in all cases.

The stabilized version, *stb*-IPTW performs rather well, even sometimes beating out the db-TMLE at the larger sample sizes, and for one parameter at $n = 100$. The advantage of TMLE in an SRCT is that under correct specification of Q (or $\bar{Q}$) it is semi-parametric efficient, and the simulations bear this out. Using the Superlearner, without the correct model in the library, the performance is slightly degraded from that under correct specification, though still good relative to the comparison estimators. It's performance using Superlearner would improve with more algorithms in the library. Examination of the EIC can tell us under what circumstances an efficient estimator (one whose influence curve spans the EIC) should have lower variance than an inefficient estimator, such as IPTW. In this particular case, that means $Pa(L(j))$ being highly predictive of $L(j)$. One could thus construct generating mechanisms that varied in their relative strengths of that prediction (see, e.g., Chaffee and van der Laan, 2012), and observe the relative efficiencies of the TMLE and *stb*-IPTW estimator.

The data-adaptive G-computation estimator occasionally performs well but, for example in estimating EY_1, it exhibited a bias that diminishes very slowly with sample size. It typically loses in terms of MSE to even the db-TMLE which is a good indication that fluctuation updating along a least favorable

submodel reduces the bias in the expected way. It seems to do well at each sample size in estimating EY_2, for reasons we have not discovered. When this estimator lucks out in terms of bias, it has a fairly low relative MSE. One must keep in mind though, that the (rather serious) drawback of this estimator is its lack of an established method of consistently estimating its variance.

Chapter 4

Applying a TMLE to the Estimation of a Causal Effect in a Long Term Observational Study

In this chapter we apply the Superlearner-based nce-TMLE to a real dataset. We describe first the background of the estimation problem before moving to the technical details of the implementation of the estimator.

Two-thirds of American women 20 years of age and above are overweight, and one-third are obese (Flegal et al., 2010). Naturally, it is hypothesized that behaviors and environmental factors play a critical role in the dramatic increases in population level obesity beginning in the late 1980s. A comprehensive overview of the current literature suggests a large number of complex causes of obesity, at both social and biological levels, with the conclusion that there is not likely to be a silver bullet for reducing obesity. Rather, the focus is on understanding how interventions targeting multiple risk factors may work to reduce future population levels of obesity (McPherson et al., 2007).

One such potential target is the amount of weight gained during pregnancy: *gestational* weight gain. While a woman's not gaining enough weight can have negative impacts on the gestation of the baby, excessive weight gain has the potential to lead to longer-term weight gain and obesity for the mother. However, a surprisingly small number of observational studies have examined this, and those that have suffer from substantial limitations. Only two studies have found that excessive gestational weight gain was associated with increased body weight 15 years after the birth (Rooney et al., 2005, Linne et al., 2004). In addition, women who still retained weight six months after delivery were more likely to weigh more 15 years later (Rooney et al., 2005). A recently published study, with the longest follow-up to date, reported that in a subsample of 2055

pregnant women, excessive gestational weight gain during pregnancy significantly doubled the odds of becoming overweight and quadrupled the odds of becoming obese 21 years after delivery, after adjusting for a wide range of potentially confounding variables, including high risk pregnancy conditions like diabetes, method of delivery, physical activity and television watching during pregnancy, depression and psycho-social factors (Mamun et al., 2010).

While these several studies give some suggestion of the potential importance of gestational weight gain for long term obesity in women, they are only correlational in nature, and thus may not give us useful information for what might result if we actually intervened on the population to reduce gestational weight gain. One of the greatest limitations of these prior correlational studies was the analysis of the effects of only a single pregnancy—a reslut of limited methodological tools available for accounting for time varying confounding. This modeled situation is a substantial departure from reality for the population of women, since so many have more than one child, and thus the results apply to a relatively limited target population.

In addition, the use of data from the National Longitudinal Survey of Youth 1979 (NLSY79) allows analysis of a much longer follow-up time than prior studies, and includes as well a sample that is socio-economically and racially and ethnically diverse and representative of the United States female population generally.

4.1 Observed Data Structure and Likelihood

The dataset consisted of all women in the NLSY79 sample who had between one and four children, inclusive, as of 2008, and who had no multiple births from a single pregnancy, i.e., no twins, triplets, etc. The oldest respondent in year 2008 was 52 and the youngest was 44. The dataset we used excluded the so-called over-sampled subsets that are included in datasets downloaded from the NLSY79 site—i.e., supplemental samples added to the demographically representative cross-sectional sample that were added for sparsity concerns, such as the military subsample and the non-hispanic blacks subsample.

The dataset contained 2246 observations, each of which included baseline covariates measured in 1979, and time-varying covariates measured annually through 1986. After 1986 the measurement waves occurred every two years up through and including 2008. Although that amounts to 19 waves of measurement, this analysis used the first 17 waves only (plus outcome at 2008), since there were no births after 2004, and hence no exposure of interest after that time.

We use the following notation for the observed longitudinal data structure. $X(k),\ k = 0, 1, ...K = 16$ represents the variable or vector of variables, X

at measurement time k. $\bar{X}(k)$ signifies the history of X up through time k: $\bar{X}(k) = \{X(0), X(1), ...X(k)\}$; whereas $\bar{X}$ represents the entire history of the variable: $\bar{X} = \bar{X}(K) = \{X(0), X(1), ...X(K)\}$.

We can think of the data as consisting of i.i.d. observations O, where

$$\begin{aligned} O =& (L(0), \delta(0), \Delta_A(0), \Delta_A(0)A(0), ..., \\ & L(K), \delta(K), \Delta_A(K), \Delta_A(K)A(K), \Delta, \Delta Y), \end{aligned}$$

with the variable coding as follows

- $L(0)$: baseline covariates, consisting of *race, age at start of study, number of term pregnancies* and associated *gestational weight gain (GWG) prior to start of study*, and the following variables, also collected at each interview wave: *education level, employment status, income level, marital status, pre-pregnancy BMI, number of previous births, smoking status (if one or more pregnancies).*
- $\delta(k), (0 \leq k \leq K)$: number of pregnancies since previous interview $\in \{0, 1, 2\}$
- $\Delta_A(k), (0 \leq k \leq K)$: indicator (yes/no) that $A(k)$ is not missing
- $A(k), (0 \leq k \leq K)$: the exposure of interest (see definition below)
- $L(k), (0 < k \leq K)$: subset of those mentioned under $L(0)$: *education level, employment status, income level, marital status, pre-pregnancy BMI, number of previous births, smoking status (if one or more pregnancies), GWG for each pregnancy*, as well as $I(interview = 1)$, an indicator that the subject was interviewed in year k.
- Δ: indicator that Y is not missing
- $Y \equiv L(K+1)$: binary outcome, I($BMI \geq BMI_0$) with $BMI_0 = 30$, and BMI obtained at year 2008, regardless of age.

4.1.1 Likelihood

The likelihood of the above described data can be factorized as

$$\begin{aligned} P(O) = P(L(0)) &\prod_{k=1}^{K} P(L(k) \mid Pa(L(k))) \times \\ &\prod_{k=0}^{K} P(\delta(k) \mid Pa(\delta(k))) \times \\ &\prod_{k=0}^{K} P(\Delta_A(k) \mid Pa(\Delta_A(k))) \times \\ &\prod_{k=0}^{K} P(A(k) \mid Pa(A(k))) \times \\ &P(\Delta \mid Pa(\Delta)) P(Y \mid \Delta = 1, Pa(\Delta))^{\Delta}, \end{aligned}$$

where we use the notation $Pa(X)$ to indicate the "parents" of X, i.e., the factors in the likelihood that occur before X in time. So, e.g., $Pa(L(k))$ is the set $\left(\bar{A}(k-1), \bar{\Delta}_A(k-1), \bar{\delta}(k-1), \bar{L}(k-1)\right)$, and $Pa(A(k)) = \left(\bar{\Delta}_A(k), \bar{\delta}(k), \bar{L}(k), \bar{A}(k-1)\right)$.

As we've described in previous chapters, it's helpful to think of this likelihood as consisting of two parts, the part associated with the variables on which we do, and the ones on which we do not, want to intervene. In keeping with our earlier notation, we use g_k, $k = 0, 1, ...K$ and $Q_{L(k)}$, $k = 0, 1, ...K+1$ to denote these two parts respectively. We can thus rewrite the above likelihood as

$$P(O) = \prod_{k=0}^{K+1} Q_{L(k),\delta(k)} \prod_{k=0}^{K} g_k,$$

where

$$\begin{aligned} Q_{L(k),\delta(k)} \equiv & P[Y \mid \Delta = 1, Pa(\Delta)]^{\Delta} && \text{for } k = K+1 \\ Q_{L(k),\delta(k)} \equiv & P[\delta(k) \mid Pa(\delta(k))] P[L(k) \mid Pa(L(k))] && \text{for } k = 0, 1, ...K \\ g_k \equiv & P[\Delta \mid Pa(\Delta)] P[A(k) \mid Pa(A(k))] P[\Delta_A(k) \mid Pa(\Delta_A(k))] && \text{for } k = K \\ g_k \equiv & P[A(k) \mid Pa(A(k))] P[\Delta_A(k) \mid Pa(\Delta_A(k))] && \text{for } k = 0, 1, ...K-1 \end{aligned} \tag{4.1}$$

The conditional densities g_k are thus associated with the intervention variables $\Delta, \Delta_A(k), A(k)$.

There are many possible ways to define the exposure of interest, gestational weight gain (GWG), each of which represents a choice that is related directly to the parameter one wants to estimate. One such choice is the binary indicator

of excessive vs. not excessive weight gain. This is a convenient definition since it reduces the complexity of the analysis while allowing an easily interpretable parameter of interest. However, since GWG is not defined when there is no pregnancy, we define instead exposure for time point k as

$$A(k) = \begin{cases} 0 & \text{if } \delta(k) = 0 \\ I(GWG(k) > w_0) & \text{if } \delta(k) = 1 \\ (I(GWG_1(k) > w_0), I(GWG_2(k) > w_0)) & \text{if } \delta(k) = 2 \end{cases} \tag{4.2}$$

for an individual-defined w_0 determined to be excessive. The dichotomous categorization of the exposure is based on the most recent Institute of Medicine (IOM) recommendations on what level of weight gain is recommended for the long term health of the child and the mother, although most of the limited evidence supporting these clinical cut-points is based on attempts at optimizing offspring health (Rasmussen and Yaktine, 2009). While evidence of the optimality for long-term health weight for the mother is almost non-existent, these cut-offs represent the most policy-relevant target for which clinicians are attempting to intervene to change weight gain during pregnancy.

Excessive weight gain is defined as

- $>$ 40 pounds for women who are underweight ($<$ 18.5 BMI)
- $>$ 35 pounds for women who are normal weight (18.5-24.9 BMI)
- $>$ 25 pounds for women who are overweight (25.0-29.9 BMI)
- $>$ 20 pounds for women who are obese ($>$ 30 BMI).

Defining exposure in this way translates easily into the desired exposure intervention: having not excessive GWG for any pregnancy in a period. In keeping with our earlier notation, we define the desired intervention as the following of an *exposure* rule, d:

$$d(\delta(k)) = \begin{cases} \delta(k) & \text{if } \delta(k) = 0 \\ 0 & \text{if } \delta(k) = 1 \\ (0,0) & \text{if } \delta(k) = 2 \end{cases}$$

4.1.2 Post-Intervention Distribution

In chapter 1 we defined the G-formula as the product across all nodes, excluding intervention nodes, of the conditional distribution of each node given its parent nodes in the model, and with the values of the intervention nodes fixed according to the intervention of interest. This formula thus expresses the distribution of $\bar{L}$ and $\bar{\delta}$ under the intervention $\bar{A} = \bar{d} \equiv (d_0(\delta(0)), d_1(\bar{\delta}(1)), ..., d_K(\bar{\delta}(K)))$, $\Delta = 1, \bar{\Delta}_A = \bar{1}$. We use superscript a to denote this intervention:

$$P^a(\bar{L}, \bar{\delta}) = \prod_{k=0}^{K+1} Q^a_{L(k),\delta(k)}(\bar{L}(k), \bar{\delta}(k)) \tag{4.3}$$

where for $k = 0, 1, ...K$,

$$\begin{aligned} &Q^a_{L(k),\delta(k)}(\bar{L}(k), \bar{\delta}(k)) \equiv \\ &\qquad P(\delta(k) \mid \bar{L}(k), \bar{A}(k-1) = \bar{d}_{k-1}, \bar{\Delta}_A(k-1) = \bar{1}_{k-1}, \bar{\delta}(k-1)) \times \\ &\qquad P(L(k) \mid \bar{A}(k-1) = \bar{d}_{k-1}, \bar{\Delta}_A(k-1) = \bar{1}_{k-1}, \bar{\delta}(k-1), \bar{L}(k-1)), \end{aligned}$$

and for $k = K + 1$,

$$Q^a_{L(k),\delta(k)}(\bar{L}(k), \bar{\delta}(k)) \equiv P(Y \mid \Delta = 1, \bar{A} = \bar{d}, \bar{\Delta}_A = \bar{1}, \bar{\delta}(K), \bar{L}(K)).$$

Here, $\bar{\Delta}_A(k-1) = \bar{1}_{k-1}$ stands for the set of relations $(\Delta_A(0) = 1, \Delta_A(1) = 1, ..., \Delta_A(k-1) = 1)$. In words, the RHS of (4.3) is the product across all L and δ nodes of the probability of each node conditional on its parents, and with all intervention nodes set to the desired intervention values. As implied above, the intervention of interest here is that all exposures are not missing, outcome is not missing and exposures are set in accordance with the exposure rule d. This representation of the intervened-on distribution will be helpful in defining our parameter of interest.

4.2 Causal Model and Counterfactuals

We assume a structural causal model (SCM) and associated causal model $\mathcal{M}^{\mathcal{F}}$ as in chapter 1.

Suppose now that we are interested in the outcomes of individuals had their exposure followed the rule d. Given a particular SCM, we can write Y_d, the so-called counterfactual outcome under rule d, as the value Y would have taken on under the intervention where $\bar{A}$ is set to the value dictated by d as specified

by the SCM. Similarly, we define Y_a as the value Y would have had under the intervention $\bar{A} = \bar{d}, \Delta = 1, \bar{\Delta}_A = 1$.

With the counterfactual outcome Y_a now defined in terms of the solution to a system of structural equations given by the SCM, we define a corresponding counterfactual parameter $\Psi^F(P_{U,X})$. For the parameter of interest here, the sequential randomization assumption (SRA), $Y_a \perp A(k), \Delta_A(k) \mid Pa(A(k))$ for $k = 0, 1, ...K$, and the coarsening at random (CAR) assumption (see below) are sufficient for identification of the causal parameter $\Psi^F(P_{U,X})$ and a particular parameter of the observed data distribution $\Psi(P_{P_{U,X}})$ for some Ψ (Robins, 1986). We define the counterfactual parameter of interest as

$$\Psi^F(P_{U,X}) = \Psi_1^F(P_{U,X}) - \Psi_2^F(P_{U,X}) \equiv \ P(Y_a > y_0) - P(Y_{\Delta=1} > y_0)$$

where y_0 is a fixed value of BMI, which we have chosen as 30. This represents the difference in two probabilities: the first is the probability of being obese (under this obesity definition) had everyone in the target population followed exposure rule d, and with no missingness on either exposure or outcome. The second probability is that of being obese under the actual conditions, and had there been no missingness on the outcome. We can rewrite Ψ^F as the difference of two means by defining $\tilde{Y} = 1$ if $Y > y_0$ and 0 otherwise. Then the above parameter is equivalent to

$$E\tilde{Y}_a - E\tilde{Y}_{\Delta=1}$$

Henceforth we drop the tilde over the Y and take as our outcome the binary indicator of obesity in 2008 under the above definition, rather than the year 2008 BMI itself.

4.3 Parameter of the Observed Distribution

Under the two assumptions mentioned above,

$$\begin{aligned} EY_a - EY_{\Delta=1} &= \ EY^a - EY^{\Delta=1} \\ &= \ \Psi_1(P) - \Psi_2(P) \end{aligned} \tag{4.4}$$

where P is a distribution of the observed data O in the model $\mathcal{M}$.

The superscripts have the same meaning as in the previous section. The translation from subscripts to superscripts, valid under the SRA and CAR,

represents the assumption that the parameter of the observed data distribution is equivalent to the corresponding parameter of the underlying data-generating mechanism.

There are various ways of mathematically representing the RHS of (4.4) in terms of the measured variables. Different representations of the parameter will inspire different estimators, since the different representations suggest estimation of different parts of the true joint and/or conditional densities (see, e.g., chapter 3). As we showed in chapter 3, one particular representation involves writing the parameter as a series of iterated conditional means. A resulting plug-in estimator of the parameter would then attempt to estimate those conditional means and would require no density estimation at all. Our parameter EY^a can again be written as a series of nested conditional expectations. The inner most such conditional expectation is

$$\begin{aligned} & E\left(Y \mid \Delta = 1, \bar{A} = \bar{d}, \bar{\Delta}_A(K) = 1, \bar{\delta}, \bar{L}(K)\right) \\ =& E\left(Y \mid Pa^a(Y)\right). \end{aligned} \tag{4.5}$$

Similar to our earlier notation, $Pa^a(X(k))$ indicates the set $Pa(X(k))$ with each intervention node set to the value dictated by the specified intervention. We use the notation $\bar{Q}^a_{K+1}$ to signify this innermost conditional expectation. Similarly, define

$$\bar{Q}^a_K \equiv E\left(\bar{Q}^a_{K+1} \mid Pa^a(L(K))\right)$$

and so on for $\bar{Q}^a_k$, $k = K-1, ...0$. The parameter we want to estimate is written in terms of these nested conditional expectations where for each k, the relevant conditional expectation is taken conditional on $\bar{A}(k-1) = \bar{d}_{k-1}, \bar{\Delta}_A(k-1) = \bar{1}_{k-1}, \bar{\delta}(k-1), \bar{L}(k-1)$. We express our parameter at any stage of the nesting set:

$$\begin{aligned} \Psi_1(P) &= \Psi_1(\bar{Q}^a) \\ &= \bar{Q}^a_0 \\ &\equiv E[E\left(\bar{Q}^a_1 \mid Pa^a(L(1))\right)] \\ &\equiv E[E\left(E(\bar{Q}^a_2 \mid Pa^a(L(2)) \mid Pa^a(L(1))\right)], \end{aligned} \tag{4.6}$$

and so on, where $\bar{Q}^a \equiv \left(\bar{Q}^a_{K+1}, \bar{Q}^a_K, ..., \bar{Q}^a_0\right)$.

Note that the $\bar{Q}_k^a$ for any k is only defined if $P\left(\bar{A}(k-1) = \bar{d}_{k-1} \mid \bar{L}(k-1), \bar{\delta}(k-1)\right) > 0$ a.e., and so we require these probabilities to be strictly postive—the so-called *positivity* or *experimental treatment* assumption.

For the second parameter in (4.4), $\Psi_2(P)$, we require the coarsening at random assumption $(Y \perp \Delta \mid \bar{A}, \bar{\delta}, \bar{L})$ in order for the causal parameter to be equivalent to the relevant parameter of the observed data distribution. The identity result is thus

$$\begin{aligned} EY_{\Delta=1} &\overset{CAR}{=} EY^{\Delta=1} \\ &= \sum_{\bar{l},\bar{s},\bar{a}} E(Y \mid \Delta = 1, \bar{L} = \bar{l}, \bar{\delta} = \bar{s}, \bar{A} = \bar{a}) P(\bar{L} = \bar{l}, \bar{\delta} = \bar{s}, \bar{A} = \bar{a}) \\ &= \sum_{\bar{l},\bar{s},\bar{a}} \bar{Q}^1(\bar{l}, \bar{s}, \bar{a}) P(\bar{L} = \bar{l}, \bar{\delta} = \bar{s}, \bar{A} = \bar{a}) \\ &= \Psi_2(P). \end{aligned} \tag{4.7}$$

Our estimation problem is now defined: we seek to estimate the parameters of the above two so-called g-computation formulas. For Ψ_1 this means estimation of each of the iterated conditional expectations implied in (4.6), and for Ψ_2 the elements of the summand expressed in (4.7).

4.3.1 Censoring of the Exposure Variable

The data-gathering methods for this study allowed for respondents to miss an interview or interviews and return later to the study for subsequent interviews. For any particular interview and respondent, a number of the variables identified as potential confounders may be missing. This included the exposure variable. However, for a respondent who missed, say, interview k but returned for an interview at monitoring time $k' > k$, both the history of the pregnancy variable, $\bar{\delta}(k')$, and the history of the exposure variables, $\bar{A}(k')$, as of k' are present. This is because, even if a respondent does not answer various questions at a particular interview period, we believe they always responded to questions about pregnancies, which included gestational weight gain information.

This data-gathering process set up an interesting set of conditions regarding missingness and censoring of the exposure variable. First, note that for each time-dependent variable, including $A(k)$, if the variable was missing for a particular respondent it was coded as 0. A corresponding column of missingness indicators was generated for each such variable subject to missingness. This indicates to the regression algorithms that a missingness variable equal to 1 signifies something different from missingness equaling 0. It also allows the

algorithms to retain in their procedures observations that have missingness rather than omitting them, which would result if missing values were just coded as "NA" and with no missingness indicator.

Suppose one now proceeds to estimate $P(\Delta_A(k) = 1 \mid Pa(\Delta_A(k))$, i.e., probability of the exposure variable not missing at time k, given the past at k. Since for some observations, the variable $\Delta_A(k)$ was collected at time $k' > k$, it may be the case that events occurring after k affect the value of $\Delta_A(k)$. This means that the variable $\Delta_A(k)$ represented in the data may not be a function only of variables generated prior to time k, which means the conditional distribution mentioned above does not capture the correct causal ordering of the data-generating process. Further, this means that among observations for which an interview occurred after k, the missingness indicators for exposures before k will perfectly predict $\Delta_A(k)$—since in those cases no exposures from the start of the study to k will be missing—resulting in probability estimates of either 1 or unreasonably close to 1. This presents some interesting problems in estimation of exposure censoring (and indeed the overall parameter) if one performs estimates at each time k.

To circumvent these issues, we proposed two slightly different versions of the target parameter Ψ_1. The first is the same as that which has been described thus far, but simply redefines missingness on exposure; the second assumes no missingness on exposure.

The first version of Ψ_1, call it $\Psi_1^{[1]}$, defines $\Delta_A(k) \equiv I(int(k) = 1)$, where $int(k) = 1$ just in case there was an interview in year k. Under this definition of $\Delta_A(k)$, $A(k)$ is considered missing if there was no interview in year k, even if in fact the exposure is present in the data. The quantity $P(\Delta_A(k) = 1 \mid Pa(\Delta_A(k))$ now represents a conditional distribution that conditions on the correct set of variables, i.e., this representation in the likelihood is the correct one based on the time ordering. This parameter thus represents a slightly different intervention from that corresponding to the original definition of $\Delta_A(k)$. The interpretation of $\Psi_1^{[1]}$, i.e., $EY^{a^{[1]}}$ is the mean outcome, under the true data-generating distribution, under the intervention $a^{[1]}$, defined as no missing interviews, following exposure rule d through all years (from 1979 on), and no missing outcome.

The second version, $\Psi_1^{[2]}$, does not include exposure missingness as an intervention variable, and thus represents the mean outcome under the intervention $a^{[2]} \equiv (\bar{A} = \bar{d}, \Delta = 1)$.

The two parameters we will be estimating are thus

$$\Psi^{[1]} \equiv \Psi_1^{[1]} - \Psi_2 \equiv EY^{a^{[1]}} - EY^{\Delta=1}$$
$$\Psi^{[2]} \equiv \Psi_1^{[2]} - \Psi_2 \equiv EY^{a^{[2]}} - EY^{\Delta=1}.$$

4.4 Estimation

Naturally we seek to demonstrate the utilization of the nce-TMLE here. Comparison estimators are of little value in the analysis since we don't know the true parameter values. The remainder of the chapter is about applying the Superlearner-based nce-TMLE to this estimation problem.

4.4.1 TMLE

Estimation of ψ_1

Let P_0 be the true observed data-generating distribution. We seek to estimate the two alternate parameters $\psi_1^{[1]} \equiv \Psi_1^{[1]}(P_0)$ and $\psi_1^{[2]} \equiv \Psi_1^{[2]}(P_0)$. The major elements of the TMLEs we describe here, (aside from their underlying theory) and details of the algorithm and its implementation are presented in chapter 3. There the parameter to be estimated was somewhat similar to Ψ_1 here, though the former parameter did not include intervention on any censoring variables since censoring was not simulated.

In the following we refer to Ψ_1 (without superscript) whenever the explicative details are the same for the two alternate parameters $\Psi_1^{[1]}$ and $\Psi_1^{[2]}$, with the obvious adjustments for the respective interventions.

Obtaining initial estimates and the corresponding fluctuated updates of the nested conditional expectations implied in (4.6) are performed as described in section 3.3.2.

In this case, let $\bar{Q}^a_{K+1,n}$ be an initial estimate of $\bar{Q}^a_{K+1}$ obtained via a preferred approach. The updated estimate $\bar{Q}^{a,*}_{K+1,n}$ is, as described previously, obtained by using the initial estimator as an offset in a univariate logistic regression of Y on the covariate $c_1(K+1)$, where for our two alternate parameters,

$$
\begin{aligned}
c_1^{[1]}(K+1) =& \frac{I(\bar{A}(K) = \bar{d}_K, \Delta = 1, \bar{\Delta}_A = \bar{1})}{\prod_{k=0}^{K} g_k^{[1]}} \\
c_1^{[2]}(K+1) =& \frac{I(\bar{A}(K) = \bar{d}_K, \Delta = 1)}{\prod_{k=0}^{K} g_k^{[2]}}
\end{aligned} \tag{4.8}
$$

For all other k, $1 < k < K+1$,

$$
\begin{aligned}
c_1^{[1]}(k) =& \frac{I(\bar{A}(k-1) = \bar{d}_{k-1}, \bar{\Delta}_A(k-1) = \bar{1}_{k-1})}{\prod_{j=0}^{k-1} g_j^{[1]}} \\
c_1^{[2]}(k) =& \frac{I(\bar{A}(k-1) = \bar{d}_{k-1})}{\prod_{j=0}^{k-1} g_j^{[2]}},
\end{aligned} \tag{4.9}
$$

where g_j and g_k are similar to the definitions in section 4.1.1, but adjusted to reflect the respective intervention variables defining $\Psi_1^{[1]}$ and $\Psi_1^{[2]}$. In observational studies such as the present one, these conditional probabilities are not known and must be estimated from the data. As with the estimation of the initial $\bar{Q}_k^a$, we prefer data adaptive estimation here as well (see below).

The estimation of each successive conditional expectation starting with $\bar{Q}_{K+1,n}^a$ is updated in this manner yielding the corresponding $\bar{Q}_{k,n}^{a,*}$ for $k = K+1, K, ..., 1$. The parameter estimate is then computed as $\psi_{1,n} \equiv \Psi_1(\bar{Q}_n^{a,*}) = \bar{Q}_{0,n}^{a,*}$

Estimation of Ψ_2

We estimate $\psi_2 \equiv \Psi_2(P_0)$ using a different TMLE, one which was developed in the context of point-treatment estimation with a single intervention node. This estimator has been developed and written about extensively (see, e.g., van der Laan and Rose, 2011). Here the entire history of all variables prior to Δ can be thought of as potential confounders, and since they all occur prior to Δ there are no time-varying confounders in the causal model. In this case the TMLE algorithm requires an initial estimator of $\bar{Q}^1(W) \equiv E(Y \mid \Delta = 1, W)$ (see (4.7)), where for notational convenience we define $W \equiv (\bar{L}(K), \bar{\delta}, \bar{A})$. Notice that the exposure(s) of interest for parameter Ψ_1—the $\bar{A}$—are not intervention nodes for this parameter, and are not otherwise distinguished from the other covariates. The only intervention variable here is the missingness indicator on the outcome.

Upon obtaining the initial estimate $\bar{Q}_n^1$ of $\bar{Q}^1$, it is updated, similarly to the process described above, by running a univariate logistic regression of Y on a covariate c_2 with the logit of the initial estimator as offset, where

$$c_2 \equiv \Delta / g(\Delta \mid W),$$

and $g(\Delta \mid W) \equiv P(\Delta \mid W)$. This regression yields a maximum likelihood estimate of the coefficient in front of c_2, ϵ. The estimator can be written

$$\begin{aligned}\psi_{n,2} =& \frac{1}{n} \sum_{i=1}^{n} \bar{Q}_n^{1,*}(W_i) \\ =& \frac{1}{n} \sum_{i=1}^{n} expit[logit(\bar{Q}_n^1(W_i)) + \epsilon_n c_{2,i}]\end{aligned}$$

where $\bar{Q}_n^1(W)$ is an initial estimate of $\bar{Q}^1(W)$ and ϵ_n is the MLE of ϵ.

To obtain $Q_n^1(W)$ one performs a "regression" of Y on W among all observations with no missing outcome. We place "regression" in quotes to indicate that standard software for estimating a conditional expectation is not implied as the estimation method. Predicted values of the outcome are then obtained for each observation using the empirical model $Q_n^1(W)$.

4.5 Implementation of TMLE and Details of the Analysis

Implementation of the TMLE of ψ_1 consists in an initial estimator of $\bar{Q}_k^a(P_0)$ for each k, which means estimating

$$\begin{aligned} &E(\bar{Q}_{k+1}^{a^{[1]}} \mid \bar{L}(k), \bar{\Delta}_A(k) = \bar{1}_k, \bar{A}(k) = \bar{d}_k, \bar{\delta}(k)), && \text{for } \Psi_1^{[1]} \\ &E(\bar{Q}_{k+1}^{a^{[2]}} \mid \bar{L}(k), \bar{A}(k) = \bar{d}_k, \bar{\delta}(k)), && \text{for } \Psi_1^{[2]} \end{aligned} \tag{4.10}$$

i.e. estimating expectation of $\bar{Q}_{k+1}^a$ conditional on $\bar{L}(k)$ among observations with exposure following rule d up through at least time k, and additionally for $\Psi_1^{[1]}$, conditional on having all interviews up through k. Here we define $Y \equiv \bar{Q}_{K+2}$ for convenience. The first regression in the series (for estimating either $\psi_1^{[1]}$ or $\psi_1^{[2]}$) is conditional on $\Delta = 1$ as well.

The update step in producing $\bar{Q}_{k,n}^{a,*}$ for each k requires an estimate of g_k, which

is equivalent to estimating a set of conditional expectations since the $A(k)$ are binary.

For estimation of $\psi_1^{[2]}$, the only observations that were included in the estimation of the parameter were those with complete treatment history. The details of the data gathering process mentioned in section 4.3.1 imply that $\Delta_A(K) = 1$ implies $\bar{\Delta}_A(K-1) = 1$, i.e., any observation with $\Delta_A(K) = 1$ has a complete treatment history. To compensate for the fact that our estimator only uses observations that meet this condition, all the regressions for estimating $\bar{Q}_k^a$ and g_k were weighted regressions, with weight

$$\frac{\Delta_A(K)}{P(\Delta_A(K) \mid Pa(\Delta_A(K)))} \tag{4.11}$$

The univariate logistic regression performed in the fluctuation update was weighted in the same way. We thus refer to this as an inverse probability of censoring weighted, reduced data TMLE (IPCW-R-TMLE), with the weights as specified above. This estimator was first proposed and developed in van der Laan (2008).

4.5.1 Super Learner

One has a number of choices in estimating these conditional expectations. One could propose a parametric model and simply do logistic regression of $\bar{Q}_{k+1}$ on the specified covariates according to the parametric model proposed. Parametric analyses are common when a coefficient in front of the exposure or treatment variable in the model can be interpreted as the parameter of interest, even when there is no evidence that the proposed model is correct. Here however, since we have defined our parameter non-parametrically, there is even less reason than in those cases to assume a particular parametric model. The ability to interpret coefficients as the parameter of interest is of no advantage and hence parametric modeling represents an assumption that is both wrong and unneeded. In contrast to such an approach, we attempt to estimate these conditional means as nonparametrically as possible. Our preferred approach is to employ the highly data-adaptive Super Learner algorithm (van der Laan et al., 2007b), which was one of the choices for initial estimator in chapter 3.

Briefly, the Super Learner (SL) uses minimum cross-validated risk as the criterion for building a regression model, with squared error as loss function. There may be other valid loss functions appropriate for a given parameter of interest—we gave another example in section 3.3.2. Cross-validated risk has many advantages as an algorithm selection criterion. Among other things,

it targets the quantity we're interested in "honestly." That is, we obtain an assessment of each algorithm's predictive success based on a validation set that was not used in construction of the prediction model generated by the algorithm.

One first determines the algorithms to be contained in the SL's "library of learners," which can include any data-adaptive algorithms suitable to the type of data at hand, as well as the user's preferred parametric regression models if there are such. The algorithm then computes an average cross-validated risk for each learner across V cross-validation folds where V is a number selected by the user. The final prediction model consists of a convex combination of the predictions from the original set of learners, each of which is weighted based partly, but not entirely, on its cross-validated risk; the lower the risk the more weight the learner will tend to have, though there are additional factors in determining the final set of weights. Indeed, some algorithms might receive a weight of 0.

The Super Learner is our preferred algorithm for estimation of conditional expectations, and TMLE requires an estimate both of $\bar{Q}_k$ for $k = 1, ..., K+1$ and g_k, $k = 0, ..., K$. We thus used SL to estimate both of these entities. The library for each regression consisted of a subset of the following algorithms (the relevant R package name is given in parentheses): main terms logistic regression (glm), Random Forests (randomForest), Bayesian Generalized Linear Models (bayesglm), generalized additive models (gam), Feed-forward Neural Networks (nnet), multivariate adaptive polynomial spline regression (polymars) and Step-wise selection based on AIC (stepAIC). The Super Learner is capable of supporting many, many more algorithms for samples of the size for this study, but time constraints required limiting the library to just a few algorithms.

4.5.2 Estimating g

In $\Psi^{[2]}$ the intervention only involves $A(k)$ and Δ, not $\Delta_A(k)$, thus g is now represented by the conditional distributions of these intervention nodes.

Recall that when $\delta(k) = 2$, the exposure variable $A(k)$ is a vector of two binary GWG variables. Let us define $A(k)$ as in (4.2) and $A(k)_1 \equiv I(GWG_1(k) > w_0)$, $A(k)_2 \equiv I(GWG_2(k) > w_0)$. We model the probability of exposure having followed d up through each k based on

$$P(A(k) = d_k \mid \bar{\delta}(k), \bar{L}(k), \bar{A}(k-1) = \bar{d}_{k-1}) \equiv$$

$$\begin{cases} 1 & \text{if } \delta(k) = 0 \\ P(A(k)_1 = d_{1,k} \mid \delta(k) = 1, Pa^{a^{[2]}}(\delta(k))) & \text{if } \delta(k) = 1 \\ P(A(k)_2 = d_{2,k} \mid A(k)_1 = d_{1,k}, \delta(k) = 2, Pa^{a^{[2]}}(\delta(k))) & \\ \quad \times P(A(k)_1 = d_{1,k} \mid \delta(k) > 0, Pa^{a^{[2]}}(\delta(k))) & \text{if } \delta(k) = 2 \end{cases}$$

As is intuitive from the definition of d, $d_{j,k} = 0$ for $j = 1, 2$ and all k. For $\Psi_1^{[2]}$ this definition encompasses g_k, though as mentioned above, all "regressions" were weighted regressions with weights given in (4.11). This means that estimation of g (and $\bar{Q}$ for that matter) using the super learner required that each learner in the library was capable of performing weighted regressions. This limited the library to four or five learners.

Estimation of the probability of following d and no exposure censoring, for the purposes of estimating $\psi_1^{[1]}$ required estimating the joint distribution of $\Delta_A(k)$ and $A(k)$, conditional on the past, for each $k = 0, 1, ...K$, as suggested in (4.1). We estimate the relevant probabilities here based on the representation above, but modified to include the time-dependent censoring on $A(k)$:

$$\begin{aligned} g_k^{[1]} \equiv P(A(k) = d_k \mid \Delta_A(k) = 1, Pa^{a^{[1]}}(\Delta_A(k))) \\ \times P(\Delta_A(k) = 1 \mid Pa^{a^{[1]}}(\Delta_A(k))) \end{aligned}$$

4.5.3 Births Prior to 1979

Some respondents had births prior to the start of the study. Since we could not measure time-dependent confounders for such observations, or all relevant pre-birth variables if just one such pregnancy, we treated the number of pre-1979 births and their associated GWGs as baseline covariates. The implication for the interpretation of our parameters is as follows: we consider the target population to be a mix of ages ranging from 14 - 22 at the beginning of a data generation process in which intervention starts at that point. Some of these subjects will have had term-pregnancies prior to that and others not. Our parameter targets this population with non-intervened on pregnancies prior to this point, and intervened-on pregnancies after this point. The target population at 2008 naturally consists of a mix of ages as well, but, as is common in clinical trials, we define the outcome at this fixed time.

4.5.4 Sparsity Issues

Another issue that arose in this data is sparsity concerning the exposure variable under certain conditions. Table 4.1 shows the numbers of women who had one and two pregnancies during each of the between-interview periods. For those who had one pregnancy, estimation of $P(A(k)_1 = d_{1,k} \mid \delta(k) = 1, Pa^a(\delta(k)))$ for year 2004 clearly presents a problem. The number of covariates in $\bar{L}(k), \bar{\delta}(k)$ for that year is well over 300. One can easily see how, even assuming this conditional probability is a smooth function of all of the covariates, it is impossible to use the data to do a reasonable job of estimating it. Other years have similar problems, though not as severe.

yr	$\delta(k) = 1$	$A(k)_1 = d(k)$	$\delta(k) = 2$	$A(k)_2 = d(k)$
1979	106	64	0	NA
1980	160	105	0	NA
1981	217	133	0	NA
1982	237	158	1	0
1983	254	157	0	NA
1984	251	162	0	NA
1985	273	183	0	NA
1986	292	246	0	NA
1988	525	352	11	8
1990	500	275	21	12
1992	448	243	17	6
1994	390	242	7	6
1996	281	196	3	3
1998	219	143	10	6
2000	135	83	3	2
2002	67	41	0	NA
2004	8	6	0	NA

Table 4.1: Numbers of observations with one pregnancy and with two pregnancies in the specified interview period, and associated number of observations which followed rule d (not excessive GWG). The total number of observations in the data set is 2246.

The problem is even worse for estimating $P(A(k)_2 = d_{2,k} \mid A(k)_1 = d_{1,k}, \delta(k) = 2, Pa^a(\delta(k)))$ in most years. The problem is particularly acute for years 2000, 1996, 1994 and 1982, in which the number of women who had two term-pregnancies and who followed d was either 1 or 0. Parametric modeling without cross-validation would result in a complete overfit of these probabilities. And for the years mentioned above, even algorithms that employ cross-validation encounter a problem, since in every fold there will fail to be at least one observation in both the training and validation sets that satisfy $A(k)_2 = d_{2,k}$. Some algorithms require any outcome value for which a probability is predicted (i.e. outcomes in the validation set) to be present in the training sample as well, and they will fail to run otherwise. On the other hand, predictions from

algorithms that do not require this condition should not be trusted since these algorithms must do an unacceptable extrapolation.

The problem is somewhat mitigated by the fact that the fraction of women having two pregnancies in a period is quite small, and therefore the estimate of our parameter may be somewhat insensitive to bias in the estimation of these probabilities. For many of the years, the condition $\delta(k) = 2$ did not arise at all in the dataset and we can ignore the issue of estimating $P(A(k)_2 \mid ...)$ in those years. Certainly in the target population there is a small, positive probability that $\delta(k) = 2$ for each k, but fortunately we can infer that this constitutes a very small fraction of the target population. The associated portions of the population will contribute very little to the true value of the parameter.

These constitute what one might call *practical* positivity violations: there is no reason to think that the relevant true conditional probabilities of exposure are particularly small, though the sparsity in this data set makes those probabilities difficult to estimate.

We dealt with these issues in two ways. First we implemented a variable reduction technique based on a Markov condition. Here, this amounts to the assumption that the probability of following d, given the entire past including baseline covariates, is equal to the probability of following d given the recent past, including baseline covariates, i.e., that the time-varying covariates from the more distant past affect the current exposure probability only through the most recent period's covariates:

$$\begin{aligned} P(A(k)_1 =& d_{j,k} \mid \delta(k) = 1, Pa^a(\delta(k))) \\ &= P(A(k)_1 = d_{1,k} \mid \delta(k) = 1, L(k), \bar{A}(k-1) = \bar{d}_{k-1}), \end{aligned}$$

and similarly for $P(A(k)_2 = d_{2,k} \mid A(k)_1 = d_{1,k}, \delta(k) = 2, Pa^a(\delta(k)))$. We did, however, include covariates in $L(k)$, such as parity, that are summary measures of earlier time-varying covariates, since these may indeed be important time-varying confounders, and predictors of the current probabilities. Invoking this Markov assumption allowed the number of conditioned-on variables to be reduced to around 20, which is a manageable number for many of the years. With this reduced set of covariates, cross-validation in the Super Learner algorithm appeared to do a reasonable job of avoiding overfitting for the years in which the number of observations with one or two pregnancies was relatively small.

For years with very small numbers of cases of $\delta = 1$ or $\delta = 2$, we adopted the following additional smoothing approach. Reasoning that the conditional distribution of $A(k)$ for a given k ought to be very close to that for year $k+1$ or $k-1$, or possibly interview periods even more removed from k, we estimated $P(A(k')_j = d_{j,k'} \mid ...)$ for some k' close to k in which the number of observations

with $\delta(k') = 1$ (or 2 as the case may be) was larger. Then, to estimate the relevant probability for each of the small number of observations in year k, we used the model for the estimated conditional probability from period k' but set the covariate values accordingly, i.e., to those of each relevant observation from year k.

4.6 Results and Discussion

The TMLEs for the various parameters, including confidence intervals are given in table 4.2.

Parameter	Estimate	95% CI	80% CI	70% CI
$\psi_1^{[1]}$	0.282	-	-	-
$\psi_1^{[2]}$	0.297	-	-	-
ψ_2	0.333	-	-	-
$\psi^{[1]} \equiv \psi_1^{[1]} - \psi_2$	-0.051	[-0.099, -0.003]	[-0.082, -0.020]	[-0.077, -0.026]
$\psi^{[2]} \equiv \psi_1^{[2]} - \psi_2$	-0.037	[-0.109, 0.034]	[-0.084, 0.009]	[-0.075, 0.001]

Table 4.2: Parameter estimates and various confidence intervals. Confidence intervals were computed only for the overall parameters of interest. (See Appendix D).

The interpretation for $\psi^{[1]}$ is a reduction of about 18% in the probability of being obese in the target population had all pregnancies been intervened on such that the GWG was not excessive, had all subjects received an interview in all years and had their outcomes not been censored, compared to the probability of being obese in the target population under the actual GWG and interview conditions, had outcomes not been censored. The effect is just significant at the 95% level.

For $\Psi^{[2]}$, the interpretation is that there is about a 12% reduction in probability of being obese had all pregnancies been not excessive in GWG and had outcomes not been censored, compared to the associated probability if outcomes had not been censored. The effect is not significant at the 95% level, but is so at about the 70% level.

It may indeed be the case that a parameter that represents an intervention in which respondents always come in for interviews ($\Psi_1^{[1]}$) is a substantially different parameter than one defined in terms merely of the presence or absence of data. Thus one should view with caution the notion that these two "versions" of the parameter really are getting at the same effect.

The variance of the estimates of ψ_n were estimated using the so-called influence curve or influence function of the TMLE (see Appendix D). In our estimation procedure we truncated both $g_{n,A(k)}$ and $\bar{Q}^a_{n,(k)}$ at the value 0.015 for all k. This is to keep the influence curve for any observation well-bounded, which in

turn bounds the variance of the estimator. There appear to be no observations in which the bounding on g_n changed the estimated value of that probability, but there were some cases where the truncated estimate of $\bar{Q}_k^a$ was different from the non-truncated estimate. Even so this truncation should produce a minimum of bias in the estimates of ψ.

One might have expected the variance of $\psi_n^{[2]}$ to be smaller than $\psi_n^{[1]}$ since the latter represents a more highly constrained intervention, i.e., fewer people follow the intervention of all pregnancies being not excessive GWG and coming in for all interviews. However, as mentioned above the estimation of $\psi^{[2]}$ required a series of weighted regressions. We noticed in this process that the predictive success for each of these weighted regressions was markedly less (in terms of MSE) than those of the unweighted regressions done for estimation of $\psi^{[1]}$. Since the influence curve of the estimator includes factors that are closely related to the MSE, one would expect a higher MSE to translate to larger variance of the IC. We believe this is a plausible explanation for the higher variance of the second estimator.

4.7 Conclusions

We have implemented a TMLE that we examined in earlier chapters, and is semi-parametric efficient under the right conditions and asymptotically unbiased (under milder conditions) to estimate the population level effect of an intervention on women's gestational weight gain on long term obesity. To our knowledge, no causal analysis of this effect has been performed to date, let alone with such an advanced estimator.

The two versions of the parameter we estimate essentially compare the probability of being obese in middle age if we could intervene on the target population such that no one had excessive weight gain during any pregnancy (and data were not missing) with the probability in the actual population of being obese later in middle age under no such intervention (and if there were no censoring on the data). Though the two parameters $\Psi^{[1]}$ and $\Psi^{[2]}$ are technically different because, among other things, they define censoring differently, they are both versions of what we (somewhat imprecisely) paraphrase above. Defining censoring in terms of having or not having an interview represents a tight constraint, and is not part of the intervention we would ideally want to perform, though it did result in an estimate with a slightly lower variance than that for the alternate version of the parameter.

Based on the risk difference point estimates of the effect, our findings suggest that the development of interventions to successfully reduce excessive gestational weight gain to within the IOM weight gain recommendations could have a substantial impact on reducing population obesity in middle age, a condition that has been linked to increased risk of mortality in other studies. The

confidence intervals, however, suggest a fairly wide range of probable impacts of intervening on the target population, from no impact at all, to a very substantial impact (i.e. 30% reduction in obesity). Thus any interventions must proceed with the acknowledgement that it is within the confidence bounds of the estimate that such interventions would have very little benefit towards reducing obesity in the popluation. However, it is also important to note that, based on our estimates, we do not anticipate any potential negative impact of these interventions. Thus, in terms of the comparative cost-effectiveness of other interventions to reduce obesity, our results suggest pursuit of interventions to target reductions in excessive weight gain during pregnancy shows promise.

Bibliography

O. Bembom and M.J. van der Laan. Statistical methods for analyzing sequentially randomized trials. *Journal of the National Cancer Institute*, 99(21): 1577–1582, 2007.

O. Bembom and M.J. van der Laan. Data-adaptive selection of the truncation level for inverse-probability-of-treatment-weighted estimators. *Berkeley Division of Biostatistics Working Paper Series. Working Paper 230*, 2008.

Paul Chaffee and Mark J. van der Laan. Discussion of evaluation of viable dynamic treatment regimes in a sequentially randomized trial of advanced prostate cancer, by Wang et al. 2012. *Journal of the American Statistical Association*, 107(498):513–517, 2012.

J. D. Faires and R. Burden. *Numerical Methods*. Thomson Brooks/Cole, Pacific Grove, CA, 3rd edition, 2003.

K. Flegal, M.D. Carroll, C.L. Ogden, and L.R. Curtin. Prevalence and trends in obesity among US adults, 1999-2008. *Journal of the American Medical Association*, 303(3):235–241, 2010.

Susan Gruber and Mark J. van der Laan. A targeted maximum likelihood estimator of a causal effect on a bounded continuous outcome. *The International Journal of Biostatistics*, 6(1), 2010.

X. Guo and A. Tsiatis. A weighted risk set estimator for survival distributions in two-stage randomization designs with censored survival data. *International Journal of Biostatistics*, 1(1), 2005.

E. Laber, M. Qian, D. Lizotte, and S.A. Murphy. Statistical inference in dynamic treatment regimes. *Revision of Univ. of Michigan, Statistics Department Technical Report 506*, 2009.

Yvonne Linne, Louise Dye, Britta Barkeling, and Stephan Rossner. Long-

term weight development in women: a 15-year follow-up of the effects of pregnancy. *Obesity Research*, 12(7):1166–78, 2004.

J.K. Lunceford, M. Davidian, and A.A. Tsiatis. Estimation of survival distributions of treatment policies in two-stage randomization designs in clinical trials. *Biometrics*, 58(1):48–57, 2002.

Abdullah A Mamun, Mansey Kinarivala, Michael J O'Callaghan, Gail M Williams, Jake M Najman, and Leonie K Callaway. Associations of excess weight gain during pregnancy with long-term maternal overweight and obesity: evidence from 21 y postpartum follow-up. *American Journal of Clinical Nutrition*, 91(5):1336–41, 2010.

K. McPherson, T. Marsh, M. Brown, and G. Britain. *Tackling obesities: future choices: Modelling future trends in obesity and the impact on health*. Department of Innovation, Universities and Skills, 2007.

S. Miyahara and A.S. Wahed. Weighted Kaplan-Meier estimators for two-stage treatment regimes. *Statistics in Medicine*, 29(25):2581–2591, 2010.

E.E.M. Moodie, R.W. Platt, and M.S. Kramer. Estimating response-maximized decision rules with applications to breastfeeding. *Journal of the American Statistical Association*, 104(485):155–165, 2009.

S.A. Murphy, M.J. van der Laan, and J.M. Robins. Marginal mean models for dynamic regimes. *Journal of the American Statistical Association*, 6: 1410–1423, 2001.

L Orellana, A Rotnitzky, and J Robins. Dynamic regime marginal structural mean models for estimation of optimal dynamic treatment regimes. *International Journal of Biostatistics*, 6(2), 2010.

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

M.L. Petersen, Y. Wang, M.J. van der Laan, D. Guzman, E. Riley, and D.R. Bangsberg. Pillbox organizers are associated with improved adherence to HIV antiretroviral therapy and viral suppression: a marginal structural model analysis. *Clinical Infectious Diseases*, 45(7):908–15, 2007.

Kristin E. Porter, Susan Gruber, Mark J. van der Laan, and Jasjeet S. Sekhon. The relative performance of targeted maximum likelihood estimators. *International Journal of Biostatistics*, 7(1), 2011.

Kathleen M. Rasmussen and Ann L. Yaktine. *Weight Gain During Pregnancy: Reexamining the Guidelines*. Committee to Reexamine IOM Pregnancy Weight Guidelines; Institute of Medicine; National Research Council; The National Academies Press, 2009. ISBN 9780309131131. URL `http://www.nap.edu/openbook.php?record_id=12584`.

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 Modeling*, 7:1393–1512, 1986.

B.L. Rooney, C.W. Schauberger, and M.A. Mathiason. Impact of perinatal weight change on long-term obesity and obesity-related illnesses. *Obstetrics and Gynecology*, 106(6):1349–56, 2005.

Ori M. Stitelman and Mark J. van der Laan. Collaborative targeted maximum likelihood for time to event data. *The International Journal of Biostatistics*, 6(1), 2010.

Ori M. Stitelman, Victor De Gruttola, and Mark J. van der Laan. A general implementation of TMLE for longitudinal data applied to causal inference in survival analysis. *U.C. Berkeley Division of Biostatistics Working Paper Series*, Working Paper 281, April 2011.

S.L. Tunis, D.E. Faries, and et. al Nyhuis, A.W. Cost-effectiveness of olanzapine as first-line treatment for schizophrenia: results from a randomize, open-label, 1-year trial. *Value Health*, 9:77–89, 2006.

Mark J. van der Laan. The construction and analysis of adaptive group sequential designs. *U.C. Berkeley Division of Biostatistics Working Paper Series*, Working Paper 232, March 2008.

Mark J. van der Laan. Targeted maximum likelihood based causal inference: Part I. *The International Journal of Biostatistics*, 6(2), 2010a.

Mark J. van der Laan. Targeted maximum likelihood based causal inference: Part II. *The International Journal of Biostatistics*, 6(2), 2010b.

Mark J. van der Laan and Susan Gruber. Targeted minimum loss based estimation of causal effects of multiple time point interventions. *The International Journal of Biostatistics*, 8(1), 2012.

Mark J. van der Laan, Eric C. Polley, and Alan E. Hubbard. Super learner. *U.C. Berkeley Division of Biostatistics Working Paper Series*, Working paper 222, 2007a. http://www.bepress.com/ucbbiostat/paper222.

Mark J. van der Laan, Eric C. Polley, and Alan E. Hubbard. Super learner. *Statistical Applications in Genetics and Molecular Biology*, 6(1), 2007b.

M.J. van der Laan and M. Petersen. Causal effect models for realistic individualized treatment and intention to treat rules. *The 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 Verlag, New York, 2003.

M.J. van der Laan and S. Rose. *Targeted Learning: Causal Inference for*

Observational and Experimental Data. Springer, Berlin Heidelberg New York, 2011.

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. Rose, and S. Gruber. Readings in targeted maximum likelihood estimation. *U.C. Berkeley Division of Biostatistics Working Paper Series*, Available at: http://works.bepress.com/sgruber/6, 2009.

A.S. Wahed and A.A. Tsiatis. Optimal estimator for the survival distribution and related quantities for treatment policies in two-stage randomization designs in clinical trials. *Biometrics*, 60(1):124–133, 2004.

A.S. Wahed and A.A. Tsiatis. Semi-parametric efficient estimation of the survival distribution for treatment policies in two-stage randomization designs in clinical trials with censored data. *Biometrika*, 60(1):147–161, 2006.

Lu Wang, Andrea Rotnitzky, Xihong Lin, Randall E. Millikan, and Peter F. Thall. Evaluation of viable dynamic treatment regimes in a sequentially randomized trial of advanced prostate cancer. *Journal of the American Statistical Association*, forthcoming, 2012. doi: 10.1080/01621459.2011.641416.

Appendix A

Efficient Influence Curve for Discrete $L(1)$

In the following, $D^*_{1,m}$ indicates the efficient influence curve for the m^{th} binary indicator of $L(1), m = 0, 1, 2, 3$, and $Pa(L(1)) = (L(0), A(0))$. We have

$$\begin{aligned} D^*_{1,0}(O) = \frac{I(A(0) = d_0(L(0)))}{g(d_0(L(0)) \mid X)} \{ & E(Y_d \mid L(1) = 0, Pa(L(1))) - \\ & \sum_{m>0} E\left[Y_d \mid L(1) = m, Pa(L(1))\right] P(L(1) = m \mid L(1) > 0, Pa(L(1)))\} \times \\ & \{I(L(1) = 0) - I(L(1) \geq 0) E[I(L(1) = 0) \mid Pa(L(1))]\}, \end{aligned}$$

where, e.g.,

$$\begin{aligned} & P(L(1) = 2 \mid L(1) > 0, Pa(L(1))) \\ & = \frac{P(L(1) = 2, L(1) > 0 \mid Pa(L(1)))}{P(L(1) > 0 \mid Pa(L(1)))} \\ & = \frac{P(L(1) = 2) \mid Pa(L(1))}{1 - P(L(1) = 0 \mid Pa(L(1)))} \\ & = \frac{P(L(1) = 2 \mid L(1) \geq 2, Pa(L(1)))}{1 - P(L(1) = 1 \mid Pa(L(1))} \times \\ & \qquad \prod_{s<2} \left[1 - P(L(1) = s \mid L(1) \geq s, Pa(L(1)))\right] \\ & = P(L(1) = 2 \mid L(1) \geq 2, Pa(L(1))) \left[1 - P(L(1) = 1 \mid L(1) \geq 1, Pa(L(1)))\right], \end{aligned}$$

and

$$\begin{aligned}
&P(L(1)=3 \mid L(1)>0, Pa(L(1))) \\
&\quad = P(L(1)=3 \mid L(1)\geq 3, Pa(L(1)))\times \\
&\qquad\qquad \prod_{s=1}^{2}\left[1-P(L(1)=s \mid L(1)\geq s, Pa(L(1)))\right] \\
&\quad = 1 * \prod_{s=1}^{2}\left[1-P(L(1)=s \mid L(1)\geq s, Pa(L(1)))\right].
\end{aligned}$$

Similarly,

$$\begin{aligned}
D^*_{1,1}(O) =& \frac{I(A(0)=d_0(L(0)))}{g(d_0(L(0)) \mid X)}\{E(Y_d \mid L(1)=1, Pa(L(1)))- \\
&\sum_{m>1} E\left[Y_d \mid L(1)=m, Pa(L(1))\right] P(L(1)=m \mid L(1)>1, Pa(L(1)))\}\times \\
&\{I(L(1)=1)-I(L(1)\geq 1)E[I(L(1)=1) \mid L(1)\geq 1, Pa(L(1))]\},
\end{aligned}$$

and

$$E[I(L(1)=m) \mid L(1)\geq m, Pa(L(1))] \equiv P(L(1)=m \mid L(1)\geq m, Pa(L(1))).$$

$D^*_{1,2}(O)$ is similar, but $D^*_{1,3}(O)=0$ since

$$\begin{aligned}
I(L(1)=3)&-I(L(1)\geq 3)E[I(L(1)=3) \mid L(1)\geq 3, Pa(L(1))] \\
&= I(L(1)=3) - I(L(1)=3) * E[I(L(1)=3) \mid L(1)\geq 3, Pa(L(1))] \\
&= I(L(1)=3) - I(L(1)=3) * P[L(1)=3 \mid L(1)\geq 3, Pa(L(1))] \\
&= I(L(1)=3) - I(L(1)=3) * 1 = 0.
\end{aligned}$$

Thus the efficient influence curve for EY_d is

$$D^*(O) = D^*_0(O) + \sum_{m=0}^{3} D^*_{1,m}(O) + D^*_2(O),$$

with $D^*_0(O)$ and $D^*_2(O)$ exactly as given in Theorem 1.

The expression for clever covariate $C_{L(1,m)}$ follows immediately from $D^*_{1,m}$ as

simply the IPCW term times the first bracketed term. So, for example, $C_{L(1,2)}$ would be

$$C_{L(1,2)} = \frac{I(A(0) = d_0(L(0)))}{g(d_0(L(0)) \mid X)} \{E(Y_d \mid L(1) = 2, Pa(L(1))) - \sum_{m>2} E\left[Y_d \mid L(1) = m, Pa(L(1))\right] P(L(1) = m \mid L(1) > 2, Pa(L(1)))\}.$$

Appendix B

Data Generation for Chapters 1 & 2 Simulations

In this appendix we describe the data generation process for each of the variables in the causal model. There are notable differences in the two major sets of simulations (i.e., the binary $L(1)$ case vs. the discrete $L(1)$ case).

- *L(0)*
  For both binary and discrete $L(1)$ cases, $L(0)$ consisted of four baseline covariates, $L(0) = (W_1, ..., W_4)^T$, three of which were distributed Normally, i.e.,

  $$(W_1, W_2, W_3)^T \sim N(\mu, \Sigma),$$

  with $\mu = (0, -0.35, 0)^T$ and with all off-diagonal terms of Σ set to 0. The fourth baseline covariate W_4 was distributed as a truncated normal, also independent of the other baseline variables. Specifically, let random variable $Z \sim N(5, 1.5^2)$. Then

  $$W_4 = \begin{cases} Z & \text{if } 2 < Z < 8 \\ 0 & \text{otherwise} \end{cases}$$

- *A(0)*
  $A(0)$ was assigned randomly for all simulations, $A(0) \sim Ber(0.5)$
- *L(1)*

  – (1) *Binary* In the binary $L(1)$ case,
  $L(1) \sim Ber([1 + exp(-(Logit[Q_{L(1)}]))]^{-1})$, where

  $$Logit[Q_{L(1)}] = \tfrac{1}{2.5}(2 - W_1 - W_4 - 2W_2^2 + 1.8W_3^2 - 3W_4W_3 + 3A(0) +$$

$2(1 - A(0)))$.

and with $W_1, ...W_4$ as defined above.

- (2) *Discrete* The conditional probabilities for each factor $L(1, m)$, $m = 0, 1, 2$, were generated as follows.

$$logit[Q_{L(1,0)}] = \tfrac{1}{6.5}[-15 - W_1 - W_4 - 2W_2^2 + 1.8W_3^2 - 3W_4W_3 + 3A(0) + 2(1 - A(0))],$$

$$logit[Q_{L(1,1)}] = logit[Q_{L(1,0)}] + 2.8,$$

$$logit[Q_{L(1,2)}] = logit[Q_{L(1,1)}] + 4.2.$$

- *A(1)*

  - *(1) Binary L(1)* $A(1)$ was set according to

$$A(1) = \begin{cases} A(0) & \text{if } L(1) = 1 \\ A(0) & \text{with probability 0.5 otherwise} \end{cases}$$

  - *(2) Discrete L(1)* $A(1)$ in the discrete case was set according to

$$A(1) = \begin{cases} A(0) & \text{if } L(1) > 1 \\ A(0) & \text{with probability 0.5 otherwise} \end{cases}$$

- *L(2)*

  - *(1) Binary L(1)* For the binary $L(1)$ simulations, $L(2) \sim Ber([1 + exp(-(Logit[Q_{L(2)}]))]^{-1})$, where

$$Logit[Q_{L(2)}] = \tfrac{1}{2.5}(2 - W_1 - W_4 - 2W_2^2 + 1.8W_3^2 - 3W_4W_3 + 3A(0) + 2(1 - A(0)) + 2L(1) - 1.5(1 - L(1)) + 6 * I(d(\bar{L}) = 1) - 6.5 * I(d(\bar{L}) = 2) - W_1(1 - A(0)) + W_4A(1))).$$

  - *(2) Discrete L(1)* For the simulations with discrete L(1), $L(2) \sim Ber([1 + exp(-(Logit[Q_{L(2)}]))]^{-1})$, where

$$Logit[Q_{L(2)}] = \tfrac{1}{6}(-7 - W_1 - W_4 - 0.7W_2^2 + 0.6W_3^2 - W_4W_3 + 9A(0) + 3(1 - A(0))) + 1.4L(1) - W_1(1 - A(0)) + W_4A(1) + 6 * I(d(\bar{L}) = 3).$$

In the above expressions $I(d(\bar{L}) = j), j = 1, 2, 3$ is equal to 1 if rule j was followed at both treatment time points (as described in section 1.4.1) and 0 otherwise.

Appendix C

Data Generation for Chapter 3 Simulations

We describe the data generation process for each of the variables in the model. Note that the logits of $Q_{L(1)}$ and $Q_{L(2)}$ are highly non-linear functions of the parents of $L(1)$ and $L(2)$, respectively.

- *L(0)* consisted of two baseline covariates, $L(0) = (W_1, W_2)^T$, where

  $$W_1 \sim 20 * \chi^2_2,$$

  and the second baseline covariate W_2 was distributed as a truncated normal. Specifically, let random variable $Z \sim N(5 + (W_1/150), 1.5^2)$. Then

  $$W_2 = \begin{cases} Z & \text{if } 2 < Z < 8 \\ 0 & \text{otherwise} \end{cases}$$

- *A(0)* was assigned randomly for all simulations, $A(0) \sim Ber(0.5)$
- *L(1)* was a single discrete-valued random variable with four levels: 0,1,2,3. Define

  $$logit\big(Q_{L(1),basis}\big) \equiv -3.5 + A0 + 2 * expit(W_1/80).$$

  Then the conditional probabilities for each factor $L(1, m)$, $m = 0, 1, 2$, were generated according to the logits

  $$\begin{aligned} logit[Q_{L(1,0)}] &= logit\big(Q_{L(1),basis}\big) + 12 * I(W_2 < 3.5), \\ logit[Q_{L(1,1)}] &= logit\big(Q_{L(1),basis}\big) + 12 * I(3.5 \leq W_2 < 5), \\ logit[Q_{L(1,2)}] &= logit\big(Q_{L(1),basis}\big) + 12 * I(5 \leq W_2 < 6.5), \end{aligned}$$

and, of course $Q_{L(1,3)} \equiv P(L(1) = 3 \mid L(1) > 2, Pa(L(1)) = 1$.

- *A(1)* was set according to

$$A(1) = \begin{cases} A(0) & \text{if } L(1) > 1 \\ A(0) & \text{with probability 0.5 otherwise.} \end{cases}$$

- $Y \equiv$*L(2)* was continuous on [0,1] with value $expit\big(logit[Q_{L(2)}]\big)$, where

$$logit\big(Q_{L(2)}\big) = \big(-2 - W_2 + 1.5A(0) + 30 * expit\big(0.65 + 0.8(L(1) - 1.3)^2\big) - 9A(1) + Z_{L(2)}\big)/6$$

with

$$Z_{L(2)} \sim N(0, 1)$$

Appendix D

Computation of Variance of the Estimators in Chapter 4

Under regularity conditions, TMLE is an asymptotically linear estimator whose influence curve (IC) is equal to the so-called efficient influence curve (EIC)(van der Laan and Rubin, 2006). One can estimate the variance of the estimator by estimating the variance of the EIC, $D^*_\psi(P_0) = D^*(Q_0, g_0, \psi)$ where P_0 is the true distribution of the observed data.

Regarding Ψ_1, Theorem 1 of van der Laan and Gruber (2012) tells us that the EIC can be written in terms of $\bar{Q}^a$ and g. We adapt the result there to our data structure and parameters of interest. For parameter $\Psi_1^{[1]}$,

$$D^*_{\psi_1^{[1]}}(\bar{Q}^{a^{[1]}}, g^{[1]}) = \sum_{k=0}^{K+1} D^*_{\psi_{1,k}^{[1]}}$$

where $g^{[1]}$ is the set $(g_k^{[1]} : k = 0, ..., K)$ and

$$D^*_{\psi_{1,K+1}^{[1]}} = c_1^{[1]}(K+1)(Y - \bar{Q}_{K+1}^{a^{[1]}})$$

and for $k = 1, ...K$,

$$D^*_{\psi_{1,k}^{[1]}} = c_1^{[1]}(k)(\bar{Q}_{k+1}^{a^{[1]}} - \bar{Q}_k^{a^{[1]}}).$$

Finally,

$$D^*_{\psi_{1,0}^{[1]}} = \bar{Q}_1^{a^{[1]}} - \Psi(\bar{Q}^{a^{[1]}}).$$

$D^*_{\psi_1^{[1]}}$ is a function of the true value of the parameter, but also a function of the true distribution of the data through $\bar{Q}^{a^{[1]}}$ and $g^{[1]}$, i.e., $D^*_{\psi_1^{[1]}} = D^*_{\psi_1^{[1]}}(P_0) =$

$D^*_{\psi_1^{[1]}}(\bar{Q}^{a^{[1]}}(P_0), g^{[1]}(P_0))$. Though we cannot know P_0, we follow the common practice of using our best estimates of those parameters in estimating the variance. Thus the TMLE estimates of $\bar{Q}^{a^{[1]}}$ and $\psi_1^{[1]}$, and the SL estimate of $g_k^{[1]}$ for all k were used in place of these true parameters in estimating the IC of the TMLE, and hence its variance. As mentioned, under regularity conditions on $\bar{Q}$ and g, the IC of the TMLE is equal to the EIC, i.e.,

$$IC_{\psi_1^{[1]}} = D^*_{\psi_1^{[1]}}(\bar{Q}^{a^{[1]}}, g^{[1]}).$$

Further, even if $\bar{Q}_n$ and g_n do not converge to $\bar{Q}$ and g, respectively, if $\bar{Q}_n$ converges to some $\bar{Q}'$ and g_n converges to some g', then the variance estimate using this estimation method is asymptotically correct, and one obtains asymptotically nominal coverage. Table 1.3 shows that estimating the TMLE variance by estimating the variance of its influence curve (under both correct and incorrect model specification) can give slightly non-conservative estimates of confidence intervals at smaller sample sizes, but they rapidly approach the nominal levels by sample sizes of around 500.

For estimation of the alternative parameter $\psi_1^{[2]}$, the influence curve is modified as a result of the inverse weighting given by (4.11), and is not equivalent to the EIC for that parameter. In this case the influence curve follows exactly the expressions given above, with the corresponding elements for intervention $a^{[2]}$, but multiplied by the weighting function given by (4.11). That is, the influence curve for our estimator of parameter $\psi_1^{[2]}$ is

$$IC_{\psi_1^{[2]}} = \left(\frac{\Delta_A(K)}{P\big(\Delta_A(K) \mid Pa(\Delta_A(K))\big)} \right) \left(D^*_{\psi_1^{[2]}} \left(\bar{Q}^{a^{[2]}}, g^{[2]} \right) \right)$$

where $D^*_{\psi_1^{[2]}}(\bar{Q}^{a^{[2]}}, g^{[2]})$ corresponds exactly to the version presented for $\Psi_1^{[1]}$ but with all the obvious adjustments based on the alternate intervention.

The EIC for parameter Ψ_2 can be written

$$D^*_{\psi_2}\left(\bar{Q}, g_2, \psi_2\right) = \frac{\Delta}{P(\Delta \mid W)} \left(Y - \bar{Q}(\Delta, W)\right) + \bar{Q}^1(W) - \psi_2,$$

with $g_2 \equiv P(\Delta \mid W)$. Here too, the IC of the TMLE is equal to the EIC under regularity conditions, i.e., $IC_{\psi_2} = D^*_{\psi_2}\left(\bar{Q}, g_2, \psi_2\right)$. As with ψ_1, the TMLE update of $\bar{Q}_k$ and the TMLE of ψ_2 and the SL estimate of g_2 are used in place of the respective parts of the EIC in the computation. And, under assumptions similar to those above, the variance estimate for the TMLE for this parameter will also have asymptotically correct coverage.

From the fact that the IC of the sum of two asymptotically linear estimators is the sum of their ICs, we can write the IC of the TMLE of ψ (supressing the intervention-related superscripts) as

$$IC_{\psi} = IC_{\psi_1} - IC_{\psi_2},$$

from which it follows, by the central limit theorem and the fact that D^* is a mean-0 function of O, that a reasonable estimate of the variance of the TMLE of ψ is

$$\widehat{var}\left(\psi_n\right) = \frac{\frac{1}{n}\sum_i \left(IC_{\psi,n}(O_i)\right)^2}{n},$$

where $IC_{\psi,n}(O_i)$ indicates the sample estimate of $IC_{\psi}(O_i)$.