Causal Inference using Targeted Maximum Likelihood Estimation

Duc Hieu Do
STUDENT NUMBER: 2039710

THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE IN DATA SCIENCE & SOCIETY
DEPARTMENT OF COGNITIVE SCIENCE & ARTIFICIAL INTELLIGENCE
SCHOOL OF HUMANITIES AND DIGITAL SCIENCES
TILBURG UNIVERSITY

Thesis committee:

Dr. Richard Starmans
Dr. Michal Klincewicz

Tilburg University
School of Humanities and Digital Sciences
Department of Cognitive Science & Artificial Intelligence
Tilburg, The Netherlands
January 2021

Acknowledgements

First of all, I would like to express my heartfelt gratitude to Dr Richard Starmans, who has always been a reliable and outstanding advisor throughout the process of conducting this thesis. The completion of this dissertation would not have been possible without his wise counsel, constructive arguments and criticisms.

Second, I want to express my love toward my parents and sister, who have always trusted and listened to me unconditionally. I wouldn't have been courageous enough to return to academia without their ongoing encouragement. There are no words to describe how grateful I am for having them by my side.

Third, I would like to thank my best friend, Ms Phuong Thao Tran. Thank you for your forgiveness, understanding, and for being with me in the darkest moments.

Finally, to the readers, I hope you will enjoy reading my thesis.

Duc Hieu Do.

TABLE OF CONTENTS

Causal Inference using Targeted Maximum Likelihood Estimation

Duc Hieu Do

In causal inference, one of the main interest is to estimate the marginal effect of the treatment on the outcome of interest. In the observational study, this process is much more challenging than that in randomized controlled trial and it require statistician or data scientists to handle the confounding covariates correctly. Many traditional confounding adjustment methods, such as G-computation or propensity score, rely on the assumption of having correct parametric models of the data-generating distribution. If these models are misspecified, the following estimates will be biased. Alternatively, Targeted Maximum Likelihood Estimation (TMLE) is a doubly robust substitution estimation method that reduces the misspecified parametric model's bias using data-adaptive machine learning algorithms while preserving statistical inference. This thesis reviewed the recent literature on TMLE and investigated the conceptual similarities and dissimilarities between TMLE and other traditional approaches. These characteristics were also highlighted by applying TMLE procedure to examine the causal relationship between smoking and the risk of having a low birth weight baby. The result indicated that, smoking during pregnancy would increase the mother's risk of having a low birth weight baby. Moreover, a comparative simulation study under realistic settings was also applied to compare these methods' performances. In general, the findings obtained are consistent with recent literature on causal inference and TMLE, highlighting the advantages of TMLE over the traditional approaches, with the lowest percent bias in almost all scenarios.

1. Introduction

Causal inference centrally focuses on detecting the causal relationship between the intervention/treatment and the outcome of interest. Nowadays, this topic is gaining significantly more attention, especially in economics, medicine, or epidemiology. One might want to research the effect of an advertising campaign on a company's profitability or the effectiveness of the cure on the treatments status. This statistical inference procedure can provide a valuable reference for a company's strategic business planning or doctors' devising treatment regiments.

The estimation of the casual effect depends significantly on which type of study being investigated. In Randomized clinical trials (RCTs), treatment is assigned randomly to ensure that the difference in the treatment before and after receive the treatment is due to the treatment effect itself (Faraoni and Schaefer, 2016). On the other hand, in observational or non-randomized studies, the way unit are assigned with specific treatment may also depends on other covariates (Austin, 2011; Faraoni and Schaefer, 2016). These covariates can also be called as confounder or confounding

covariates, if they also affect the outcome. Therefore, if statisticians want to estimate the marginal treatment effect alone, they must carefully adjust for these confounding covariates.

There are two most traditional methods that researchers often use to estimate causal effects while adjusting for the confounding covariates. The first method is called G-computation, which is an outcome regression-based method. Specifically, G-computation adjusts for confounder via a parametric regression model of the outcome mechanism that incorporates the treatment and controls for all measured covariates. Meanwhile, the Propensity Score (PS)-based methods are adopted using the probability of a person being assigned to a particular intervention/treatment given the baseline confounding covariates (Austin, 2011; Elze et al., 2017; Rosenbaum and Rubin, 1983). Despite being relied on two different parametric models of the treatment and outcome mechanism, both propensity score-based method and G-computation method will get biased estimates if their models are misspecified (Austin, 2011; Schuler and Rose, 2017). Alternatively, Double Robustness (DR) is a method that combines the two aforementioned methods together. Therefore, this method has a double chance to receive an unbiased effect estimator if either the parametric regression model for the outcome or for the treatment mechanism is correctly specified (Słoczyński and Wooldridge, 2018). Targeted Maximum Likelihood Estimation or TMLE, introduced by (Van Der Laan and Rubin, 2006), is a doubly robust, maximum-likelihood–estimation method that applied statistical learning and machine learning to flexibly adjust for confounding covariates while making minimal functional form assumption of both the treatment and the outcome (Van Der Laan and Sherri Rose, 2011). Moreover, TMLE integrates a second "targeting" phase to reduce bias for the estimation of the parameter of interest (Schuler and Rose, 2017).

The motivation of this thesis is to discuss and shed light on the TMLE procedure for causal inference in the observational study. Specifically, based on reviewing literature with simulation and experimental study, the following two main research questions will be answered in this thesis:

***RQ1**: What makes TMLE different and why should statisticians and data scientists use this procedure instead of commonly used methods to estimate causal effect?*

To answer this research question, this thesis will discuss and investigate the conceptual similarities and differences between TMLE over traditional methods; thereby, highlight the advantages of TMLE. Later in this thesis, a step by step procedure of TMLE will also be implemented to demonstrate those similarities and differences in practice. Specifically, TMLE procedure will be applied to quantify the causal effect of smoking during pregnancy on the risk of having a low baby's born weight. However, it is also vital to validate and examine insights retrieved from theory and formulas; thus, leading to the second research question that will be discussed in this thesis is:

***RQ2**: In what way and to what extent could TMLE be integrated in practice? Compare the performance of TMLE versus that of its predecessors.*

To answer this research question, this thesis will implement a comparative simulation study to compare the performance of TMLE over other traditional methods. Five different scenarios will be set up to evaluate the extent of TMLE's application.

The remainder part of this thesis can be structured as follows. In the next section, this dissertation will begin by presenting some fundamental causal inference background. Next, four key methods to estimate causal effect in this thesis will be introduced, including G-computation, IPW, A-IPW, and TMLE. In section three, this dissertation will apply TMLE to quantify the causal effect between smoking during pregnancy and the risk of giving birth to a low birth weight baby. This section aims to introduce to the reader a basic process of TMLE in practice, thereby highlighting the differences and similarities between TMLE and other algorithms. In section four, the thesis will compare the aforementioned methods through simulation to compare each method's performance. Via this simulation, this thesis will analyze investigate which method outperforms under specific settings and scenarios. This thesis will be concluded by interpreting the results of the experimental study and the comparative simulation.

2. Background and Related Work

2.1 Notations and Potential Outcomes framework

Defining the observed data consists of n independent and identically distributed observations of $O = \{W, A, Y\} \sim P_0$, where W represents the vector of baseline covariates, A stands for the binary treatment assignment indicator, and Y denotes the outcome of interest. The P_0 denotes the true probability distribution of O. Therefore, for a set of values of (W_i, A_i, Y_i); $P_0(W_i, A_i, Y_i)$ accounts for the probability of that W, A, Y is equal to W_i, A_i, Y_i. From now on, for notational convenience, any estimation with subscript 0 will represent the truth, while the subscript n stands for an estimation under n observations.

To detect the causal effect of a particular treatment, the statisticians and data scientists can estimate the difference between the outcomes that would have occurred under each of the treatment assignments. This approach is formulated based on the Neyman-Rubin potential outcome framework. Specifically, given the observed data O, each observation i has a pair of potential outcomes $Y(1)$ and $Y(0)$, which can be interpreted as the outcome for that unit if it is treated and untreated, respectively. Consequently, the treatment effect on that single unit can be defined to be $Y(1)$ - $Y(0)$. Alternatively, one common parameter of interest in causal inference is the Average Treatment Effect (ATE), defined as follows:

$$ATE = E[Y(1) - Y(0)]$$

which can be interpreted as the the average effect of moving an entire population from exposed to unexposed group.

However, one of the fundamental problem in causal inference is that, statisticians and scientists can only be able to assign one and only one treatment or intervention to a specific unit and observe the outcome from there (Holland, 1986). For example, given a person receives the treatment A_1, then the outcome $Y(1)$ will be observed while $Y(0)$ becomes a counterfactual - an unobserved outcome that would have occurred if that person did not get the treatment beforehand. This absence of counterfactual outcomes makes estimating treatment effect become a missing data problem. There are different methods to tackle this problem and identify ATE, depending on the mechanism of how treatment is allocated.

In Randomized clinical trials (RCTs), treatment is randomly between individuals to ensure that units in both treatment and control groups share identical underlying characteristics. By this, the only factor that can contribute to differences in the outcome is due to the treatment effect itself (Faraoni and Schaefer, 2016). Therefore, the unbiased estimation of ATE can be estimated by a direct comparison of the outcome between the treatment groups (Austin, 2011). Conversely, in observational or non-randomized studies, the way a unit is assigned to which treatment mechanism often differs in that unit's characteristics; thus, affect the outcome (Schuler and Rose, 2017). Failure to isolate the effect of those confounding covariates, namely those variables associate with both the treatment and the outcome, can lead to a biased estimation of the treatment effect (Austin, 2011; Schuler and Rose, 2017).

2.2 Causal Assumptions

In order to identify ATE in observational studies, three assumptions below must be held:

Assumption 1: Stable Unit Treatment Value Assumption (SUTVA)
The first assumption is the central assumption under the Rubin causal framework (Holland, 1986). This assumption is a combination of two sub-assumptions: "No interference" and "Homogeneity of treatment version," respectively. The first sub-assumption assumes that the effectiveness of the treatment on one unit is independent of one another's treatment assignment. Given an example of violation as follows: if unit A's level of diabetes depends on whether unit B's is treated; then, the treatment effect on A will be based on the treatment assignment of not only A but also B. If then, statisticians and data scientists would need to define various potential outcomes for one unit not just with each treatment assignment assigned to that unit but also with each combination of treatment assignments received by other units. Therefore, the assumption of "No interference" should be held to quantify, specifically the treatment effect against the control. The second sub-assumption of SUTVA ensures that the same treatment version is applied to all units who received that treatment in the sample. This assumption enables the statisticians and data scientists to evaluate the treatment effect subjectively without caring about the bias due to differences in treatment level among units in that sample.

Assumption 2: Ignorability

$$Y(1), Y(0) \perp A|W$$

This assumption is sometimes referred to as no unmeasured confounders. When it is held, an observational study can be considered as an RCTs study. Specifically, in this assumption, statisticians and data scientists assume that the exposure mechanism and potential outcomes are independent, given the set of measured covariates. (i.e., no unmeasured confounding and no selection bias)

Assumption 3: Positivity

$$0 < P_0(A = a|W) < 1$$

It is assumed said that every individual in the population will have nonzero probability of adopting all kinds of the treatment, being conditioned on the baseline covariates. The intuition of this assumption is that statisticians and data scientists must be able

to collect data where they can learn about what would happen under either treatment assignment, because if there is any subgroup that has no possibility of receiving that treatment, it would be impossible to compute the ATE.

When all three assumptions 1, 2, and 3 hold, ATE can be identified. Recall that, ATE can be formulated as $E[Y(1) - Y(0)]$. Mathematically, there is nothing to change if one conditioned $E[Y(1)]$ and average over the distribution of W; thus, $E[Y(1)]$ can also be represented as $E_W\left(Y(1) \mid W\right)$. Next, when the Positivity and Ignorability assumptions are held, $E\left(Y(1) \mid W\right)$ can also be represented as $E\left[E\left(Y(1) \mid A = 1, W\right)\right.$, since there exists probability of $A = 1$ (Positivity) and $Y(1)$ is conditionally dependent on A and W (Ignorability). Finally, under the SUTVA assumption that the outcome Y is equal to the potential outcome, given the treatment assignment and the baseline covariates; thus, $E\left[E\left(Y(1) \mid A = 1, W\right)\right.$ is also equal to $E[E(Y \mid A = 1, W)$.

In summary, under the three Identify assumptions of causal effect, the average treatment effect can be represented as follows:

$$\begin{aligned} E\left[Y(1) - Y(0)\right] &= E_W\left[E\left(Y(1) \mid W\right) - E\left(Y(0) \mid W\right)\right] \\ &= E_W\left[E\left(Y(1) \mid A = 1, W\right) - E\left(Y(0) \mid A = 0, W\right)\right] \\ &= E_W[E(Y \mid A = 1, W) - E(Y \mid A = 0, W)] \end{aligned}$$

Therefore, the ATE can now be presented as the marginal effect of the treatment, adjusted for all baseline confounding covariates.

2.3 Relevant work

Under the potential outcome framework by Rubin (2005), there are basically three different main groups of technique to estimate the marginal effect of the treatment, or, in other words, the difference between the potential outcome value and the predicted counterfactual outcome value. These three methods are parametric, non-parametric and double robustness.

For the parametric estimation methods, statisticians and data scientists often rely on a parametric functional form of the outcome as a function of the exposure and other baseline covariates. The parameter of interest can then be quantified by conditioning for all of the explanatory variables, except for the exposure or computation of the difference between the model's outcomes when adjusting for the treatment assignment. Some standard parametric methods to estimate the causal effect would be the outcome regression or G-computation (Gelman and Hill, 2006; Snowden, Rose, and Mortimer, 2011; Naimi, Cole, and Kennedy, 2017). However, these methods under parametric assumptions are frequently and justifiably getting concerns about the problem of misspecified statistical models that would lead to biased estimation of the treatment effect. It is unlikely that statisticians and data scientists could know the true functional form of the data generating process in common practice, especially when dealing with high-dimensionality data (Van Der Laan and Sherri Rose, 2011). Some researchers have introduced different methods that apply data-adaptive methods like random forest (Wager and Athey, 2018) to flexibly learn the functional form while relieving the parametric assumptions. Alternatively, Ho et al. (2007) proposed a pre-processing method on input data via matching procedures to make the causal effect estimation through parametric assumptions less sensitive to the arbitrary choice of model specification.

For non-parametric methods, statisticians and data scientists often quantify the marginal treatment effect through the help of propensity score (Rosenbaum and Rubin, 1983; Austin, 2011). The propensity score is the likelihood that an individual adopting a particular treatment given the baseline covariates. Simply speaking, the statisticians and data scientists reduce the original dimension of the baseline covariates into a single new score to balance the effect from those variables on units among treatment groups. The difference of this propensity score is then suggested that it can be represented for the difference in all the potential covariates, under the ignorability assumption. Two standard and most frequently used methods using propensity score are Propensity score Matching (PSM) and Inverse Probability Weighting (IPW). Despite their widespread use in the causal inference field, some critics have risen about the efficiency and performance of these methods (King and Nielsen, 2019; Allan et al., 2020). Moreover, although both matching and weighting are non-parametric, the propensity score estimation falls under parametric assumptions of a functional form relationships between the treatment assignment and the observed covariates. Therefore, the efficiency of PSM and IPW depend on the correct specification of the propensity score model. Similar to the case of G-computation, researchers also tried to overcome this weakness using data adaptive machine learning methods (Lee, Lessler, and Stuart, 2010; Ferri-García and Del Mar Rueda, 2020).

The Doubly Robust (DR) methods can apply both non-parametric and parametric methods to adjust the treatment effect for covariates (Bang and Robins, 2005). As its name must imply, these methods have a double opportunity to efficiently estimate the parameter of interest if either the model for the outcome regression or the model for the treatment assignment mechanism (i.e., the propensity score) is correctly specified (Funk et al., 2011; Han, 2018). A fundamental doubly robust estimator is Augmented Inverse Propensity Weighting (AIPW) (Robins, Rotnitzky, and Zhao, 1994). Nevertheless, in their comprehensive research, Kang and Schafer (2007) criticized the performance of DR methods against conventional methods using regression-based prediction or re-weighting with propensity score, especially when both models are incorrectly specified.

Applying data-adaptive methods in estimating treatment effect has raised many interests in the causal inference society. Specifically, in the observational study, when researchers hardly know the true data-generating distribution, relying on a single conventional linear or logistic regression may lead to inefficiency (Snowden, Rose, and Mortimer, 2011). Moreover, in the era of big-data like nowadays, where high-dimensional data sets with hundreds or even thousands of potential covariates have become more popular; it would be impossible to specify the parametric statistical model for the problem correctly. Instead, data-adaptive machine learning approaches automatically select variables and construct the functional form based on the data itself to reduce the bias of misspecified parametric models (Chernozhukov et al., 2016; McConnell and Lindner, 2019; Künzel et al., 2019). However, data-adaptive algorithms, for instance, regression tree is not always better than the standard parametric model like logistic regression (Austin, Tu, and Lee, 2010). There is an alternative procedure named super learner (Laan and Dudoit, 2003; Van Der Laan, Polley, and Hubbard, 2007; Polley and Laan, 2010). Super Learner is an ensemble algorithm which contains a library of estimators that may include not only various traditional regressions but also data adaptive-based approaches. This algorithm uses Cross-Validation to train and

validate the winning candidate that can best minimize the user-specified loss functions (i.e., Mean Squared Error or Mean Absolute Error) of the prediction from the candidate. Therefore, the super learner is guaranteed to perform better than any procedure by simply including that procedure in the library (van der Laan and Starmans, 2014). Alternatively, instead of selecting a discrete winner with the most optimal prediction bias, the super learner can select a weighted combination of various algorithms that is also optimized to minimize the specified loss function. This weighted combination is investigated that, in many realistic scenarios, outperform the next best algorithm, both in term of variance and bias (Van Der Laan and Sherri Rose, 2011; Luque-Fernandez et al., 2018).

One might apply the winning algorithm or a set of weighted algorithms from super learner procedure as replacements for the parametric models of the treatment and outcome mechanisms and estimate the ATE based on that. However, in their study, Van Der Laan and Sherri Rose (2011) introduced two problems that an estimator may encounter if statisticians and data scientists naively apply data-adaptive algorithms such as super learner procedure. Firstly, the main prioritization of super learner is to maximize the prediction/classification function of the outcome, given the treatment and all measured covariates. However, this also means that while it could estimate the whole conditional density of Y, it would also including a portion of density that is not under statisticians and data scientists main interest (i.e., the target estimand). For example, in logistic regression, a maximum likelihood estimation aims to estimate the coefficients of all parameters, while statisticians and data scientists only care about the coefficient in front of the treatment. Therefore, direct plug-in estimator based on the super learner in a sample with size n may lead to biased causal effect with a convergence rate of around $n^{-1/4}$ (Mark van der Laan, 2017). Secondly, no statistical theory can be applied to analyze super learners estimators' variances estimation. Therefore, statisticians and data scientists must come up with non-parametric bootstrap, which may computationally intensive for statistical inference (Van Der Laan and Sherri Rose, 2011).

Targeted Maximum Likelihood Estimation (TMLE), introduced by Van Der Laan and Rubin (2006), is a doubly robust, efficient substitution estimator that possesses the ability to incorporate flexible, data-adaptive algorithms into inferential statistics while conserving formal statistical inference (Mark van der Laan, 2017). TMLE's characteristics make it becomes a very appealing approach to estimate the causal effect in an observational study. First, TMLE is a doubly robust procedure that is consistent if either the outcome regression or the propensity score model is consistently estimated. An estimator is consistent if its estimate converges to the true value of the parameter being estimated, as the sample size increases. Furthermore, TMLE is asymptotically efficient (i.e., converges at the fastest rate) if both outcome and treatment mechanism are consistently estimated (Van Der Laan and Sherri Rose, 2011). Second, TMLE is a substitution estimator, which is stable and robust against outliers because it respects the global constraints of the model and ensures that the estimates fall within the parameter space (Lendle, Fireman, and Van Der Laan, 2013; Schuler and Rose, 2017). Third, TMLE's targeting feature resolves the aforementioned drawbacks of incorporating super learner for the estimation of outcome and treatment mechanism (Van Der Laan and Sherri Rose, 2011).

2.4 Estimation Methods

In this section, four methods of interest in this thesis will be presented.

2.4.1 G-computation. The G-computation or Simple Substitution estimators estimate the standardized difference between the conditional mean of the outcome Y between treatment groups A, given the baseline covariates W:

$$\hat{\psi}_n^{G-comp} = \frac{1}{n} \sum_{i=1}^{n} \left[\overline{\mathrm{Q}}_n \left(1, \mathrm{W_i}\right) - \overline{\mathrm{Q}}_n \left(1, \mathrm{W_i}\right) \right] \tag{1}$$

where $\overline{\mathrm{Q}}_n \left(A, \mathrm{W}\right) \equiv \hat{E} \left(Y \mid A, W\right)$ is the conditional expectation of Y, given covariates W for those who received the treatment and for those who are not, respectively. These conditional expectation can be estimated through using standard parametric regression of model Y. For instance, given a binary outcome, a logistic regression models of the outcome, conditional on the treatment and a transpose vector W of the baseline covariates as follow:

$$logit[\mathrm{P}(\mathrm{Y} = 1 \mid \mathrm{A}, W)] = \beta_0 + \beta_1 \mathrm{A} + \beta_2^{\mathrm{T}} W$$

From there, we can estimate the conditional expectation of the outcome from the vector of the estimates for the coefficients of the model parameters:

$$\overline{\mathrm{Q}}_n \left(1, \mathrm{W}\right) = expit(\beta_0 + \beta_1 1 + \beta_2^{\mathrm{T}} W)$$
$$\overline{\mathrm{Q}}_n \left(0, \mathrm{W}\right) = expit(\beta_0 + \beta_1 0 + \beta_2^{\mathrm{T}} W)$$

where $expit$ stands for the inverse logistic function. Then, these estimates can be plugged in the equation (1) to estimate the ATE. Note that, the process of estimating the coefficient of the model parameters and apply more sophisticated methods, for instance, data adaptive machine learning or super learner.

2.4.2 Propensity Score.

Every propensity score-based method starts with the estimation of this score. The propensity score can be interpreted as the conditional likelihood of an unit being exposed, given the baseline covariates. Therefore, this score can be estimated, traditionally, via logistic regression:

$$logit[\hat{\mathrm{P}}(\mathrm{A} = 1, W)] = \alpha_0 + \alpha_1^{\mathrm{T}} W$$

Then, the propensity score can be estimated by taking the inverse of the above logistic function:

$$\hat{\mathrm{P}}(\mathrm{A} = 1, W) = expit(\alpha_0 + \alpha_1^{\mathrm{T}} W) \tag{2}$$

From now on, for notational convenience, the estimated propensity score will be denoted as $\hat{g}\left(1 \mid W\right)$

2.4.2.1 Inverse Propensity Score Weighting - IPW.

In the observational data, there will be some units, who, conditional on the baseline covariates, will be more or less likely to receive the treatment. For example, assume that statisticians and data scientists want to estimate the ATE of the medication on young people, and they observed that every 9 out of 10 individuals would receive that treatment. If they calculate the difference of the outcome between treated and untreated

groups, they will not be able to infer whether that result belongs to the causal effect of the medication or just because of "being a male".

Therefore, IPW's goal is to create a "pseudo-population" which tends to make all the covariates evenly distributed between treated and control groups. For this purpose, IPW up-weight under-presented (i.e, those with high-probability of being assigned the treatment but in the control group) treatment/control subjects and down-weight those who are over-presented (i.e, those with high probability of being assigned the treatment and really receive that treatment). Mathematically, IPW apply the inverse of the propensity score to each unit in the exposed group $\frac{1}{\hat{g}(1|W)}$ and the inverse of not being in the treated group for those in the comparator $\frac{1}{1-\hat{g}(0|W)}$. Generally, IPW estimates the ATE by computing the difference between the weighted mean of the outcome in the treated and control group, mathematically formulated as:

$$\hat{\psi}_n^{IPW} = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{A_i Y_i}{\hat{g}_n(1|W_i)} - \frac{(1-A_i)Y_i}{1-\hat{g}_n(1|W_i)}\right] \tag{3}$$

2.4.2.2 Augmented Inverse Probability Weighting - AIPW.

The augmented Inverse Probability Weighting is the method that combine outcome regression and IPW together. The first job is to specify the outcome regression model and the propensity score model. AIPW is then double robust and gives us unbiased estimates given that any of those two models are correctly specified. To do this, AIPW constructs its estimator by solving the influence curve equation. The more specific explanations to illustrate this property are presented as follow:

Recall the formula to estimate the conditional expectation if everyone in the population is treated, using IPW as:

$$\hat{E}^{IPW}\left(Y \mid A=1, W\right) = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{A_i Y_i}{\hat{g}_n(1|W_i)}\right]$$

This estimation will be biased if the model of the propensity score $\hat{g}_n\left(1 \mid W_i\right)$ is misspecified. To improve this situation, AIPW adds an "augmented term" to the above equation:

$$\hat{E}^{A-IPW}\left(Y \mid A=1, W\right) = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{A_i Y_i}{\hat{g}_n(1|W_i)} - \frac{A_i - \hat{g}_n(1|W_i)}{\hat{g}_n(1|W_i)}\overline{\mathrm{Q}}_n\left(1, \mathrm{W_i}\right)\right] \tag{4}$$

$$= \frac{1}{n}\sum_{i=1}^{n}\left[\frac{A_i(Y_i - \overline{\mathrm{Q}}^0(1, \mathrm{W_i})}{\hat{g}_n(1|W_i)} + \overline{\mathrm{Q}}_n\left(1, \mathrm{W_i}\right)\right] \tag{5}$$

First, using the equation (4) to consider the situation when statisticians and data scientists correctly specified the propensity score model but fail to specify the correct outcome regression model. By correctly specified the propensity model, it meant that the bias term between the expectation of A given W and the propensity score equals to 0. Therefore, the augmented term will equal to 0 no matter what value of $\overline{\mathrm{Q}}_n\left(1, \mathrm{W_i}\right)$ is. Next, by simply re-arranging the equation (4) to have equation (5). Considering the second case when statisticians and data scientists correctly specified the outcome regression model but the propensity score model is specified. Analogously, the bias term between the expectation of Y_i and $\overline{\mathrm{Q}}_n\left(1, \mathrm{W_i}\right)$ is equal to 0 as the outcome

regression model is correctly specified; thus, it is no matter whether the functional form of the propensity score is wrong.

The characteristic of being double robustness apply the same when estimating the conditional expectation for everyone in the control group. Generally, the A-IPW estimator of the ATE can be formulated as follow:

$$\hat{\psi}_n^{A-IPW} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{A_i}{\hat{g}_n(1|W_i)} - \frac{1-A_i}{\hat{g}_n(0|W_i)}\right)\left(Y_i - \overline{\mathrm{Q}}_n\left(A_i, \mathrm{W_i}\right)\right) + \overline{\mathrm{Q}}_n\left(1, \mathrm{W_i}\right) - \overline{\mathrm{Q}}_n\left(0, \mathrm{W_i}\right)$$

2.4.3 Targeted Maximum Likelihood Estimation - TMLE.

Targeted maximum likelihood estimation is a general methodology introduced by Van Der Laan and Rubin (2006) for constructing doubly-robust semi-parametric efficient estimators. Similar to A-IPW, the double robustness property means that the TMLE estimator is consistent if either, but not necessarily both, the outcome model or the treatment model is correctly specified. Moreover, the TMLE has minimal variance among the class of semi-parametric estimators when both models are correctly specified.

Procedurally, TMLE starts with estimate the preliminary conditional expectation of the outcome, given the treatment and the baseline covariates denoted $\overline{\mathrm{Q}}_n^0\left(A, \mathrm{W}\right)$. The superscript 0 denotes that it is the initial estimation of $E_0\left(Y \mid A, W\right)$. These estimates will then can be applied to predict the potential and counterfactual outcomes under each exposure group as in previous G-computation methodology. Traditionally, this first process is often conducted through a standard parametric regression model (G-computation). Alternatively, instead of betting on one parametric model, TMLE will adopt the super learner to non-parametrically accomplishing this process. Generally speaking, the super learner is an ensemble algorithm that includes a library of many algorithms together. Such algorithms may range from parametric models, for instance, standard linear/logistic regression to machine learning algorithms like Bayesian Regression Tree or Neural Network. After specifying the library of potential candidates of algorithms, the super learner will use Cross-validation to measure those candidate estimators' performance via the user-specified loss function, for instance, the Mean Squared Error. Finally, either the winner among potential candidates or the best-weighted combination of these potential algorithms will be selected. Super Learner will apply the winner/best-weighted combination of these estimators to predict $\overline{\mathrm{Q}}_n^0\left(A, \mathrm{W}\right)$.

One might directly use initial Super Learner estimate to predict the potential outcomes for each individual under both treated and non-treated groups, denoted as $\overline{\mathrm{Q}}_n^0\left(1, \mathrm{W}\right)$ and $\overline{\mathrm{Q}}_n^0\left(0, \mathrm{W}\right)$, respectively. However, simply plug in these estimates into the target parameter mapping will lead to biased. Recall that the Super Learner's prioritization is to maximize the prediction/classification performance of the outcome, given the treatment and the baseline covariates. In other words, Super Learner only concerns about how well the selected estimator fits the true conditional density of the outcome Y, instead of focusing on the difference in the conditional mean of Y between treatment groups as the main parameter of interest. This phenomenon depicts the wrong bias-variance trade-off for the actual causal effect of interests from using naive data-adaptive machine learning approach. Moreover, no statistical theory can be applied to analyze super learners estimators' variance estimation. Alternatively,

statisticians and data scientists must come up with non-parametric bootstrap, which may computationally intensive for statistical inference.

To break through this weakness of using data-adaptive machine learning algorithms, TMLE integrates the second "Targeting" stage to induce any residual confounding after the initial estimation. Generally speaking, this step aims to incorporate the target parameter mapping ψ to the biased initial estimator of the probability distribution of the data.

The intuition behind the "Targeting" phase of TMLE relies heavily on the definition of the efficient influence curve. Recall that, an estimator $\hat{\Psi}\left(\mathbb{P}_n\right)$ of the estimand $\Psi\left(\mathbb{P}_0\right)$ is asymptotically linear with the influence curve $IC(O_i)$ if:

$$\psi_n - \psi_0 = \frac{1}{n}\sum_{i=1}^{n} IC(O_i) + o_P\left(\frac{1}{\sqrt{n}}\right)$$

in which, the remainder $o_P\left(\frac{1}{\sqrt{n}}\right)$ will always converges to 0 in probability as the sample size goes to infinity. Therefore, if an estimators want to be efficient, it must ensure that the influence curve has finite variance with the empirical mean of the influence curve is 0, which then, can also be called as efficient influence curve. The formulation of this efficient influence curve must be derived based on given estimation problem. Specifically, the efficient influence curve for the ATE parameter can be defined as:

$$EIC\left(O_i\right) = H(A,W)\left[Y_i - \overline{Q}^*\left(A_i, W\right)\right] + \overline{Q}^*\left(1, W\right) - \overline{Q}^*\left(0, W\right) - \Psi\left(\mathbb{P}_0\right) \tag{7}$$

Noted that, this efficient influence curve TMLE using is exactly the same with the intuition of A-IPW. More sophisticated explanation and ascertainment of this formula can be found in Chapter 5 and Appendix A (Van Der Laan and Sherri Rose, 2011). For illustrative purpose, the equation (7) can be divided into two sub-equations. The first sub-equation of the equation is $H(A,W_i)\left[Y_i - \bar{Q}_0\left(A_i, W_i\right)\right]$, denoted as equation (7.1). The second sub-equation is $\bar{Q}_0\left(1, W_i\right) - \bar{Q}_0\left(0, W_i\right) - \Psi\left(\mathbb{P}_0\right)$, denoted as (7.2). Since TMLE is also a plug-in estimator like G-computation, the empirical mean of the equation (7.2) already equal to 0.

Therefore, the only concern left for TMLE to handle is to ensure that equation (7.1) also have empirical mean 0. The "Targeting" phase of TMLE will serve this purpose, by introducing the parametric model, formulated as:

$$\overline{\mathrm{Q}}_0\left(A, \mathrm{W}\right) = \overline{\mathrm{Q}}_n^0\left(A, \mathrm{W}\right) + \epsilon \mathrm{H}\left(A, \mathrm{W}\right) \tag{8}$$

where $\mathrm{H}\left(A, \mathrm{W}\right)$ is a "clever covariate" that derived from the propensity score and formulated similar to the IPW as $\left(\frac{A_i}{\hat{P}(A=1|W)} - \frac{(1-A_i)}{1-\hat{P}(A=1|W)}\right)$. This clever covariates defines which direction should be use fluctuate the initial estimator to reduce the bias to ψ_0. Meanwhile, the ϵ in the given sub-model is a fluctuation parameter that defines the magnitude of the fluctuation. This ϵ can be estimated with standard maximum likelihood estimation through a logistic regression of the outcome Y, on the clever covariate and initial fit $\overline{\mathrm{Q}}^0\left(A, \mathrm{W}\right)$ as an offset. Afterward, TMLE will be able to incorporate the information from the propensity score, through the clever covariate. To do so, TMLE update the initial estimate of $\overline{\mathrm{Q}}_n^0\left(1, \mathrm{W}\right)$ and $\overline{\mathrm{Q}}_n^0\left(0, \mathrm{W}\right)$ according to the fluctuation model:

$$\overline{\mathrm{Q}}^{*}(1,W) = expit\left[logit\left(\overline{\mathrm{Q}}^{0}(1,W)\right) + \hat{\epsilon}\hat{H}\left(1,W\right)\right]$$
$$\overline{\mathrm{Q}}^{*}(0,W) = expit\left[logit\left(\overline{\mathrm{Q}}^{0}(0,W)\right) + \hat{\epsilon}\hat{H}\left(0,W\right)\right]$$

with the superscript * denotes that it is the updated estimation. Finally, TMLE plug in the updated estimate from the previous step into the target parameter mapping:

$$\hat{\psi}_{n}^{TMLE} = \frac{1}{n}\sum_{i=1}^{n}\left[\overline{Q}_{n}^{*}\left(1,W_{i}\right) - \overline{Q}_{n}^{*}\left(0,W_{i}\right)\right]$$

3. Causal Inference of Smoking during pregnancy on the risk of having a low birth-weight baby

The initial motivation for this experimental study is illustrate the TMLE procedure and its application in causal inference with real data set. Specifically, the TMLE method is applied to evaluate the marginal causal effect if the mother smoke during pregnancy on the risk of having low birth weight baby. Besides, by analyzing and interpreting the procedure step by step, this study wants to highlight why and how TMLE is similar or different to traditional approach (G-computation, IPW, A-IPW) in practice.

3.1 The Data set

3.1.1 Description of the Data set.

The data set using for this section is the Birth Weight Risk data set, freely retrieved from the study of D. W. Hosmer and S. Lemeshow (2016), which originally conducted to investigate risk factors associated with low infant birth weight using the information of 189 women at Baystate Medical Center in Springfield, Massachusetts. This data set is analyzed in this thesis to estimate the causal effect between smoking during pregnancy of mothers against their baby's risks of having low birth-weight, using Targeted Maximum Likelihood Estimation procedure. The data set include the binary treatment variable "Smoke", with 1 is smoking at least once during pregnancy and with 0 for never having cigarettes during that time of period. The data set includes both continuous variable newborn baby's weight and categorical outcome variable that indicate whether or not the weight is low, denoted as "BWT" and "Low", respectively. This thesis will use "Low" variable as the outcome variable and interpret the causal effect later as the percentage of having low birth weight baby. Besides, for the potential covariates, denoted as W. This data set includes 8 different variables: mother's age ("Age"), mother's weights in pound ("LWT"), mother's race ("Race"), number of previous premature labours ("PTL"), history of hypertension ("HT"), presence of uterine irritability ("UI"), and number of physician visits during the first trimester ("FTV"). More sophisticated descriptions of the data will be presented in the Appendix.

3.1.2 Pre-Processing the Data set.

The data set described above can be accessed freely on the website and can be downloaded in the form of a *.csv file. However, the original format of some variables, for instance, "Age" or "LWT" are in the undesired data type (i.e., being a factor instead of an integer). Therefore, two simple functions were constructed to reformat these variables into the desired form. Besides, since there were neither missing values nor

multi-collinearity between variables, no further pre-processing step is needed. Table 1 illustrates a small sample of the data after being pre-processed:

Table 1
First five observations of the data set after being pre-processed.

	Low	Age	LWT	Race	Smoker	PTL	Hypertension	UI	FTV
1	0	19	182	2	0	0	0	1	0
2	0	33	155	3	0	0	0	0	3
3	0	20	105	1	1	0	0	0	1
4	0	21	108	1	1	0	0	1	2
5	0	18	107	1	1	0	0	1	0

3.1.3 Exploratory Data Analysis.

This section aims to retrieve some insight knowledge of the Birth Weight Risk data set. Recall that the research question behind using this data set is to estimate the causal effect of smoking during pregnancy on the risk of having a low birth weight infant. Applying the exploratory analysis to the data set, among 189 women who participated in the study, 59 cases reported having a low birth weight baby. Furthermore, among these 189 women, 74 people claimed that they had smoked at least once during the pregnancy, and 30 (40.5%) of them were reported having a low birth weight baby. Meanwhile, among 115 women who did not smoke during that period, only 29 (25.2%) persons had a low birth weight infant. Another more statistical and transparent way of interpreting this is that the phenomenon of having a low birth weight kid is more prevalent among mothers who smoked during their pregnancy than those who did not (40.2% vs. 25.2%). Next, a chi-square test of independence is applied to investigate the relationship between the treatment variable "Smoke" and the outcome variable "Low". The test result depicted that the relation between these two variables is statistically significant, X^2 (1, N = 189) = 4.923, p = .02649. Although statisticians and data scientists can say that smoking mother were more likely to have low birth weight baby, this does not imply the causal relationship between these two variables. Therefore, TMLE will be implemented in the next section.

3.2 Causal inference using TMLE

The process of estimating the causal effect using TMLE is implemented manually, using the procedure that already stated in the background section. Instead of single parametric modes for both treatment and outcome mechanism, the super learner is applied to exploit information from the data itself flexibly and adaptively. Specifically, in this study, the algorithms included in the library of the super learner library are:

- *SL.glm*: conventional main terms logistic regression
- *SL.step*: stepwise regression.
- *SL.gam*: generalised additive models.
- *SL.glm.interaction*: a logistic regression variant that includes second order polynomials and pairwise interactions of the main terms.

- *SL.ranger*: random forest as prediction method.

Note that, the choice of algorithms included in super learner library above is arbitrary and is motivated from the latter simulation results. The super learner procedure with the specified library of algorithms is applied to generate the initial estimation of the expected low birth weight risk, given the exposure (Smoke) and baseline covariates $\overline{\mathrm{Q}}_n^0\left(Smoke, \mathrm{W}\right)$. This estimation's coefficients is then used to estimate the hypothetical potential outcomes $\overline{\mathrm{Q}}_n^0\left(1, \mathrm{W}\right)$ and $\overline{\mathrm{Q}}_n^0\left(0, \mathrm{W}\right)$, for Smoker = 1 and Smoker = 0 for each individual in the data set. These potential outcomes can be interpreted as the initially expected probability of having a low birth weight baby for both treatment assignment groups and given the baseline covariates. Noted that, these estimated $\overline{\mathrm{Q}}_n^0\left(1, \mathrm{W}\right)$ and $\overline{\mathrm{Q}}_n^0\left(0, \mathrm{W}\right)$ can be immediately plugged into the target parameter mapping as equation (1) to get the G-computation's estimation for ATE. However, several drawbacks for this naive substitution, for instance, wrong bias-variance trade-off and no good approach to statistical inference are already described and explained in the previous section.

Table 2 presents the first five observations in the extended data set with initial estimation of the potential outcomes $\overline{\mathrm{Q}}_n^0\left(1, \mathrm{W}\right)$ and $\overline{\mathrm{Q}}_n^0\left(0, \mathrm{W}\right)$:

Table 2
Extended data set with initial estimation of the conditional expectation of the outcome and corresponding potential outcomes.

	Low	$Smoker$	$\overline{\mathrm{Q}}_n^0(Smoke, W)$	$\overline{\mathrm{Q}}_n^0(1, W)$	$\overline{\mathrm{Q}}_n^0(0, W)$
1	0	0	0.219	0.394	0.219
2	0	0	0.146	0.290	0.146
3	0	1	0.317	0.317	0.193
4	0	1	0.450	0.450	0.298
5	0	1	0.449	0.449	0.291

In the next step, the super learner is again applied to estimate the propensity score $\hat{g}(1, W)$. This score is generated for every observation in the data set and can be interpreted as their probability of having cigarettes during pregnancy, given baseline covariates. Similarly, the probability of not smoking is also estimated for each individual in the data set. Subsequently, the Ignorability assumption of this data set is numerically validated. Based on the statistical summary of the score's distribution, it is shown that every subgroup in the data has a positive probability of being a smoker or vice versa (min= 0.073, mean = 0.393, max = 0.942). In Figure 1 below, the Positivity assumption is also graphically inspected using the histogram that illustrates the propensity score frequency under both treatment groups (Smoke vs. no smoke). It illustrates that there is a noticeable region of overlap between the two groups, indicating a reasonable Positivity assumption

The estimated propensity score is then applied to estimate the clever covariate for each individual in the data set. Recall that, for the average treatment effect parameter, the functional form of the clever covariate is similar to the formula of the IPW that can be formulated as follows:

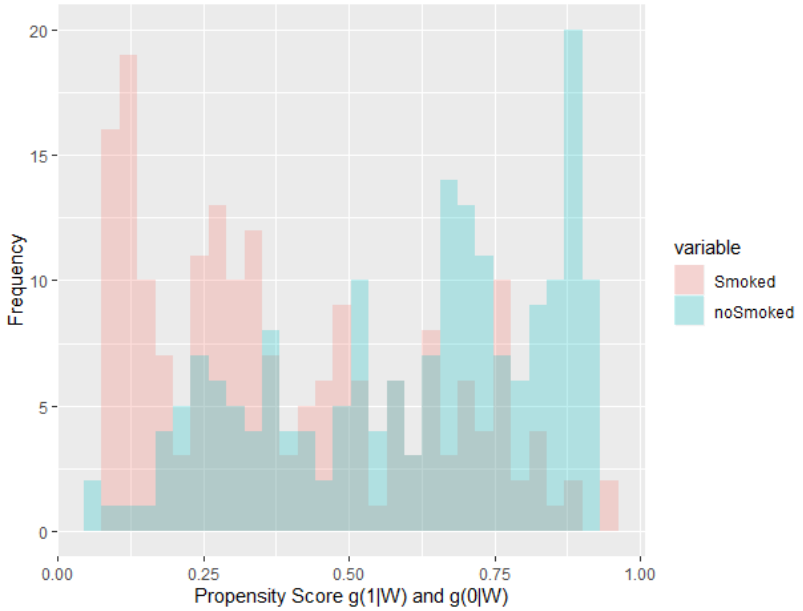

Figure 1
Overlaid histogram of propensity scores for Smoke versus noSmoke

$$\mathrm{H}\left(Smoker, \mathrm{W}\right) = \frac{Smoker_i}{\hat{g}_n(1,W)} - \frac{(1-Smoker_i)}{\hat{g}_n(0,W)}$$

In the case of IPW, this clever covariate is used to re-weight each observation's outcome based on its assigned treatment and probability of receiving that treatment. Therefore, the consistency of IPW estimator relies significantly on the correct specification of the propensity score model. Besides, the inverse-probability procedure of IPW makes it sensitive over "near" Positivity assumption violation. Meanwhile, TMLE uses this clever covariate to define a fluctuation direction that can remove the bias of the initial estimator $\overline{\mathrm{Q}}_n^0(Smoke, W)$. In addition, TMLE calculates the clever covariate at Smoker = 1 and Smoker = 0 for all observations, denoted as $H(1, W)$ and $H(0, W)$, respectively. Table 3 below illustrates the first five observations in the data set, with their according estimated propensity score and clever covariates:

Table 3
Extended data set with propensity score and corresponding clever covariates.

	Low	$Smoker$	$\hat{g}_n(1, W)$	$\hat{g}_n(0, W)$	$H(Smoke, W)$	$H(1, W)$	$H(0, W)$
1	0	0	0.249	0.750	-1.333	4.015	-1.331
2	0	0	0.162	0.837	-1.194	6.144	-1.194
3	0	1	0.651	0.348	1.533	1.533	-2.873
4	0	1	0.633	0.366	1.577	1.577	-2.731
5	0	1	0.822	0.177	1.215	1.215	-5.644

The next step is to target the initial estimator of the conditional mean outcome

$\overline{\mathrm{Q}}_n^0(A,W)$ with information in the estimated propensity score $\hat{g}_n(1,W)$. This can be done via the parametric submodel in aforementioned equation (8):

$$Low = \overline{\mathrm{Q}}_n^0\left(Smoke, \mathrm{W}\right) + \epsilon \hat{H}\left(Smoke, \mathrm{W}\right)$$

which, the parameter ϵ represents the magnitude of the fluctuation. When the of the outcome model is already correctly specified and that initial estimator $\overline{\mathrm{Q}}_n^0\left(Smoke, \mathrm{W}\right)$ has low variability with the outcome Low, ϵ will be estimated close to zero. Therefore, TMLE's next step uses maximum likelihood estimation to fit ϵ via a parametric logistic regression of outcome variable Low on the clever covariate and $\overline{\mathrm{Q}}_n^0\left(Smoke, \mathrm{W}\right)$ is held as an offset. The estimated fluctuation parameter ϵ and clever covariate for specific treatment assignment will then be used to fluctuate the initial estimator $\overline{\mathrm{Q}}_n^0\left(1, \mathrm{W}\right)$ and $\overline{\mathrm{Q}}_n^0\left(0, \mathrm{W}\right)$ by following equations:

$$\overline{\mathrm{Q}}_n^*(1,W) = expit\left[logit\left(\overline{\mathrm{Q}}_n^0(1,W)\right) + \hat{\epsilon}\hat{H}\left(1,W\right)\right]$$
$$\overline{\mathrm{Q}}_n^*(0,W) = expit\left[logit\left(\overline{\mathrm{Q}}_n^0(0,W)\right) + \hat{\epsilon}\hat{H}\left(0,W\right)\right]$$

Table 4 illustrates the final data set with the epsilon and the updated estimators $\overline{\mathrm{Q}}_{nn}^{\ 0}\left(1, \mathrm{W}\right)$ and $\overline{\mathrm{Q}}_n^0\left(0, \mathrm{W}\right)$:

Table 4
Final extended data set with updated estimators $\overline{\mathrm{Q}}^0\left(1, \mathrm{W}\right)$ and $\overline{\mathrm{Q}}^0\left(0, \mathrm{W}\right)$

	$\overline{\mathrm{Q}}_n^0(1,W)$	$\overline{\mathrm{Q}}_n^0(0,W)$	$\hat{g}_n(1,W)$	$\hat{\epsilon}$	$\overline{\mathrm{Q}}_n^*(1,W)$	$\overline{\mathrm{Q}}_n^*(0,W)$
1	0.394	0.219	0.249	0.013	0.216	0.407
2	0.290	0.146	0.162	0.013	0.144	0.308
3	0.317	0.193	0.651	0.013	0.322	0.322
4	0.450	0.298	0.633	0.013	0.455	0.455
5	0.449	0.291	0.822	0.013	0.453	0.453

The final step of TMLE is to caculate the average difference between the updated probability of having low birth weight baby when the mother smoked and not smoked during their pregnancy period, given the baseline covariates. Mathematically, the ATE using TMLE can be formulated as follows:

$$\hat{\psi} = \frac{1}{n}\sum_{i=1}^{n}\left[\overline{Q}^*\left(1,W_i\right) - \overline{Q}_n^*\left(0,W_i\right)\right]$$

3.3 Result and Inference

Both Naive TMLE - using only traditional logistic regression for the outcome and propensity score models and SL-TMLE with more powerful machine learning algorithms are applied. The ATE is estimated for both approaches step by step via the procedure described above. Besides, the influence function variance estimator, introduced in equation (7) is used to obtain statistical inference such as standard error, 95% confidence interval and p-value.

Recall the null hypothesis of this statistical problem is that, under causal assumptions, there is a causal effect between smoking during pregnancy and the risk of having low weight infant. Using naive TMLE with standard logistic regression for both outcome

and treatment mechanism, the point estimate is 0.239, with two-sided p-value = 0.018 and a 95% confidence interval of [0.040, 0.438]. This analysis suggests that there is statistically significant causal effect between smoking and the risk of having a low birth infant. Meanwhile, when more data-adaptive machine learning algorithms in super learner were applied, the SL-TMLE also suggested statistically significant casual effects of smoking. Specifically, the point estimate from SL-TMLE is 0.172 and is statistically significant, with p-value < 0.001 and a confidence interval of [0.085, 0.258]. Therefore, under causal assumptions, this results indicated that smoking during pregnancy would increase the risk of having low birth weight baby by 17.2%. Table 5 summarizes the results:

Table 5
Causal effect of Smoking on the risk of having low birth weight baby

Implementation	ATE	95% Confidence Interval	p-value
Naive TMLE	0.239	[0.040, 0.438]	0.018
SL-TMLE	0.172	[0.085, 0.258]	<0.001

4. Comparative Simulation Study

This section aims to design a comparative simulation analysis to demonstrate and compare the performances of the methods elaborated in the previous sections. Given that correcting the estimation of ATE is difficult to obtain and interpret as a baseline in the observational study; a Monte Carlo simulation is designed to evaluate the performance of G-computation, IPW, A-IPW, and TMLE under various different scenarios.

4.1 Data Generation Process

The simulation study's data is motivated and adjusted from the study example of Schuler and Rose (2017) and Luque-Fernandez et al. (2018) for estimating the causal effect of regular physical exercise on the depression status while controlling for measured confounding covariates, including gender, marital status, receipt of receiving psychological therapy, and history of using antidepressant medication. Specifically, the four covariates sex, marital status, receipt of receiving psychological therapy, and history of using antidepressant medication were generated independently and denoted as $W1, W2, W3, W4$, respectively. In which, $W1, W3$, and $W4$ were arbitrarily simulated from Bernoulli distribution with probabilities equal to 0.4, 0.5, and 0.5, respectively. Meanwhile, the relationship status variable is generated as an ordinal variable with three levels: 0 for never-married, 1 for married- spouse present, and 2 for divorce. The value for $W2$ is retrieved using a random uniform distribution with minimum zero and a maximum two and rounded off to the nearest integer. Next, the binary indicator variable A stands for the treatment assignment where $A = 1$ stands for having regular exercise, and $A = 0$ for otherwise. Similarly, the depression status variable Y is the binary variable representing whether a person is depressed ($Y = 1$) or not ($Y = 0$). The values of A and Y were generated according to Bernoulli distribution with a probability of membership retrieved from a log-linear model. The coefficients for the true functional form of the data distribution were intentionally selected so that the generated data will observe near Positivity violation. Besides, in both treatment and outcome models, an

interaction term between $W3$ and $W4$ is included to compare the performance between misspecified and correctly specified parametric models of each of the estimators. The true ATE is then computed by using the correct specified functional form of the data-generating distribution. This value is equal to -0.2249 or -22.49%, representing that, under the assumption of this artificial data set, having a regular exercise decreases the risk of getting depression by 22.49%. Note that, this statistical value has no valid statistical inference that can be generalised to the real-world causal relationship between having regular exercise and depression. Instead, the simulated ATE can only be used to evaluate estimators' performance.

4.2 Design & Procedure

This simulation aims to study to what extent may TMLE be applied and compare how it performs against the conventional methods, for instance, G-computation, IPW, and A-IPW. There are five different scenarios.

The first scenario set up situation when the parametric models for both the outcome and treatment mechanism are incorrectly specified. This scenario stands for common practice when statisticians and data scientists rely on a single linear/logistic regression model for the prediction of the conditional density of the outcome/treatment. The next two scenarios are simulated to evaluate the double robustness characteristic of TMLE, A-IPW by setting whether the outcome or treatment model is correctly specified. Besides, by omitting completely a variable from the parametric regression model, this study wants to investigate the impact of an unmeasured confounding covariate on the estimators. Next, instead of relying on a single parametric model, super learner procedure is applied to prevent bias from model misspecification. The fourth scenario is to apply the default library of super learner, which only includes three algorithms: logistic regression, stepwise regression and logistic regression with all pairwise interaction. Besides, this library of algorithms is also adjusted in the fifth scenario, with more powerful data-adaptive algorithms, for instance, generalized additive model, regression tree or random forest. The purpose behind splitting super learner library is to investigate the impact of the algorithm's choice against the performance of TMLE and other estimators. Table 6 summarizes the information of all five simulated scenarios:

Table 6
Summary of the simulated scenarios

Short Name	Scenario Description
Misspecified	Dual misspecification - standard logistic regression
Cor.Outcome	Outcome correctly specified, Treatment omitted variable
Cor.Treatment	Treatment correctly specified, Outcome omitted variable
Default-SL	Dual misspecification - default super learner
Adaptive-SL	Dual misspecification - data-adaptive super learner

In order to evaluate and compare the performance of the methods, this study reports the estimates of the mean ATE in percentage, the mean bias in percentage, and coverage confidence interval that stands for the percentage of the replications in which 95% CI contained the true value of the ATE.

4.3 Simulation Results

Table 7
Estimate of the Average Treatment Effect using G-computation, IPW, A-IPW and TMLE with a sample sized of 200

Scenario	Methods	Mean ATE	|Mean Bias %|	Coverage CI (%)
Misspecified	G-computation	-21.986	0.506	94.958
	IPW	-18.090	4.401	94.428
	A-IPW	-19.797	2.695	94.696
	TMLE	-20.364	2.128	94.652
Cor.Outcome	G-computation	-22.611	0.119	94.998
	IPW	-19.195	3.296	94.033
	A-IPW	-21.687	0.805	94.939
	TMLE	-21.235	1.257	94.841
Cor.Treatment	G-computation	-19.810	2.681	93.812
	IPW	-18.622	3.869	94.654
	A-IPW	-20.093	2.398	94.785
	TMLE	-21.222	1.270	94.879
Default-SL	G-computation	-21.304	1.187	94.805
	IPW	-20.116	2.376	94.796
	A-IPW	-20.571	1.920	94.819
	TMLE	-21.198	1.293	94.877
Adaptive-SL	G-computation	-20.915	1.577	94.634
	IPW	-17.953	4.539	94.386
	A-IPW	-20.430	2.062	94.812
	TMLE	-21.090	1.402	94.838

Table 7 compares the aforementioned methods' performance over a small sample ($n = 200$). In the first scenario, G-computation estimator outperforms other estimators, including A-IPW and TMLE, regardless of their double robust characteristic. This phenomenon may suggest the impact of the Positivity assumption's violation. Since IPW and A-IPW rely on the inverse weighting of the propensity score, they are sensitive with practical positivity violation. Meanwhile, although TMLE also uses the information from the propensity score to update the initial expectation of the potential outcomes, it does so using a logit scale. Therefore, TMLE suffers less than IPW and A-IWP on the impact of the positivity assumption. As a consequence, except for G-computation, three other methods produce anti-conservative coverage 95% CI

In the second and third scenarios, the study set up situations when the outcome or treatment model is correctly specified while the other model is significantly misspecified. In general, the results illustrate a consistency between theory and experiment over the characteristic of being double robustness of TMLE and A-IPW. However, under positivity violation and the sample size is small (n = 200), even with a correct specification of the treatment model, IPW and A-IPW still achieved bias result.

The fourth scenario is setup using the default library of algorithms in the super learner, including logistic regression on the main term, step-wise regression and logistic regression with all possible pairwise interaction. Simulation results show a significant bias reduction over standard main term regression is recognized for estimators applying propensity score. In contrast, the simple plug-in G-computation estimator based on the super learner has received high bias comparing with the result obtained through misspecified outcome model. Next, in the last scenario with powerful data-adaptive machine learning algorithm, for example, random forest and generalized additive model, the performance of all estimators are inferior to the result retrieved from the previous scenario. This phenomenon may suggest that, applying data-adaptive algorithms for the estimation of the potential outcomes and propensity score is not guaranteed for a better estimators' performance.

To evaluate how these methods behave under different sample sizes and investigate to the consistency of TMLE over other estimators. Table 8 illustrates the performance of the discussed estimators under the sample sized $n = 3000$:

Table 8
Estimate of the Average Treatment Effect using G-computation, IPW, A-IPW and TMLE with a sample sized of 3000

Scenario	Methods	Mean ATE	|Mean Bias %|	Coverage CI (%)
Misspecified	G-computation	-21.908	0.583	94.193
	IPW	-21.737	0.755	94.336
	A-IPW	-22.219	0.272	94.908
	TMLE	-22.204	0.287	94.885
Cor.Outcome	G-computation	-22.439	0.053	95.000
	IPW	-20.463	2.029	87.697
	A-IPW	-22.532	0.039	94.997
	TMLE	-22.523	0.030	94.998
Cor.Treatment	G-computation	-19.649	2.843	74.708
	IPW	-22.462	0.029	95.000
	A-IPW	-22.500	0.008	94.999
	TMLE	-20.511	0.018	94.999
Default-SL	G-computation	-22.074	0.418	94.589
	IPW	-22.407	0.085	94.990
	A-IPW	-22.430	0.062	94.995
	TMLE	-22.440	0.052	94.996
Adaptive-SL	G-computation	-21.759	0.732	93.781
	IPW	-21.417	1.079	93.734
	A-IPW	-22.220	0.291	94.889
	TMLE	-22.220	0.283	94.889

In general, with a large sample size (n=3000), almost all methods achieve noticeably better results, in term of bias, throughout all five scenarios. In the first scenario, when neither the outcome regression model nor the treatment model is correctly specified, G-computation's performance remains unchanged as the sample size increases.

Meanwhile, IPW, A-IPW and TMLE perform relatively well, with remarkable bias reduction comparing to their performance in small sample size. In the next two scenarios, both A-IPW and TMLE demonstrate their characteristics of being double robustness and consistent estimators. Specifically, if either the outcome regression model or the propensity score model is estimated consistently, A-IPW and TMLE will be consistent, regardless of near Positivity violation and omitted variable. By being consistent, these two method's estimation bias converges to 0 as the sample size increases.

When super learner with a library of default algorithms is applied, IPW, A-IPW and TMLE produce results almost as good as those obtained under the second and third scenarios with correct specification of the models. Meanwhile, although, methods using data-adaptive algorithms score lower bias as the sample size is large, their rate of convergence to the true function (bias = 0) is much slower compared to results of methods using default super learner.

Lastly, it can be noticed throughout all scenarios that, with a large sample size, A-IPW and TMLE achieve lowest bias, with coverage of CI are the closest to 95%.

5. Discussion

The ultimate motivation of this thesis is to conduct research around Targeted Maximum Likelihood Estimation and shed light on its potential over traditional approaches, including G-computation, Inverse Propensity Score Weighting (IPW) and Augmented Inverse Propensity Score Weighting (A-IPW). Two main research question were formulated to support this purpose:

***RQ1**: What make TMLE different and why should statisticians and data scientists use this procedure instead of commonly used methods to estimate causal effect?*

To answer this question, this dissertation synthesized relevant works conducted by other researchers around TMLE and conventional confounding adjustment methods such as G-computation, Inverse Propensity Score Weighting (IPW), or Augmented Inverse Propensity Score Weighting (A-IPW). The potentials of TMLE were investigated through reasoning equations behind it. In general, in comparison with G-computation and Inverse Propensity Score Weighting, TMLE has the advantage of being a doubly robust estimator. This characteristic ensures TMLE to be able to double the performance of reducing bias when estimating ATE. Besides that, with the widespread popularity of the high dimensionality data set, relying on a single parametric model for either outcome or propensity score mechanisms proved ineffective. Even with more data-adaptive machine learning algorithms or ensemble algorithms like the super learner, naively plug-in those machine learning fit into G-computation, or IPW may only leads to significant bias. Machine learning methods can flexibly and data-adaptively learn the functional form of the data-generating process by capturing convex, nonlinear relationships. Therefore, in high-dimensionality data set with thousands of possible covariates, instead of assuming on a single correct model, machine learning methods may far better fit and strengthen the validity of the assumption of having no unmeasured confounders. However, machine learning methods optimally trade-off regularization and overfitting w.r.t the whole functional density while we only caring about estimating the parameter of interest. TMLE deals with this phenomenon

effectively with its targeting step.

Comparing to A-IPW, TMLE shares some similarities and differences. First, both A-IPW and TMLE are constructed based on the efficient influence curve. Next, they are both double robustness if either the model for the treatment or outcome mechanism is correctly specified. However, the A-IPW estimator generates the point estimate by solving the efficient influence curve equation equal to 0. Therefore, in a realistic scenario with a possible violation of the Positivity assumption, A-IPW estimates may fall outside the natural bound of the problem (e.g. [0, 1]) when some observations have propensity scores close to the bound. Meanwhile, TMLE is a substitution estimator using a logit scale to update the initial fit using propensity score before plugging that updated fit into the parameter mapping. Therefore, TMLE is more stable to Positivity violation and will always respect the global constraint of the model.

Furthermore, to highlight the aforementioned characteristics of the TMLE procedure in a real case study, TMLE was applied step by step to estimate the marginal causal effect of smoking during pregnancy on the risk of having a low birth-weight baby. Observed results depict a statistically significant causal effect for both TMLE using only standard logistic regression and TMLE using super learner. The results received were intuitive with given context: smoking factor increases the risk of having low birth weight baby.

Based on the insights retrieved from theory and procedure in practice, there is no doubt that TMLE is the most appealing method to estimate the causal effect, comparing to G-computation, IPW and A-IPW. Nevertheless, regardless of the above advantages over conventional methods, TMLE should not be blindly implemented in all observational studies. It is, therefore, leading to the second research question.

***RQ2**: In what way and to what extent could TMLE be integrated in practice? Compare the performance of TMLE versus its predecessors.*

To answer the above research question, a comparative simulation study of TMLE and other confounding adjustment methods is applied. This simulation study extends the previous literature done by other researchers in various aspects. First, this study compares the relative performance of TMLE over traditional approaches, under realistic settings of both positivity violation and unmeasured confounding covariates. Second, the study considers the super learner procedure to reduce bias from the parametric outcome and treatment models' misspecification. Additionally, the effect of choosing the algorithm included in the super learner on the performance of the methods has also been assessed. Finally, the simulation study also addressed the impact of different sample sizes on the methods' performance.

Fundamentally, results from the simulation study highlighted TMLE strengths over conventional confounding adjustment methods. First, the results demonstrated the doubly robust characteristic of TMLE, which is being consistent if either the outcome model or the treatment model is correctly specified. Next, the results validated the advantage of TMLE over other propensity score-based methods like A-IPW and IPW, when there is a presence of Positivity violation, and the sample size is small. Meanwhile, TMLE did not perform better than the G-computation with a misspecified parametric regression in such situation. Next, the super learner with machine learning algorithms was set up to evaluate its performance of reducing bias from the model

misspecifcation. As the results illustrate, super learner performed better than or at least equal to parametric regression on the main terms for IPW, A-IPW and TMLE. Meanwhile, wrong bias-variance trade-off problem of super learner is confirmed with the G-computation results, as super learner only achieved equal or even larger bias than simple parametric regression. Besides, the different settings of super learner also provided insights about the impact of blindly apply powerful data-adaptive machine learning algorithms, for example, random forest or decision tree regardless of the problem. Given the observational data generated from the simulation, the adaptive super leaner performed much worse than that of the default setting of super learner, with only logistic regression, stepwise regression and logistic regression with all pairwise interaction. Even when sample size increases, the rate of convergence of data adaptive-super learner is still considerably low and no better than the parametric regression's result.

To conclude, the results retrieved from this comparative study are consistent with up-to-date literature on causal inference using TMLE. Obtained results demonstrated the double robustness, sensitivity to the violation of Positivity and Ignorability assumptions of TMLE, with the lowest percent bias in almost all scenarios. Besides, the findings indicated that, even with TMLE, incorporating data-adaptive machine learning algorithms does not guarantee to outperform the parametric regression when estimating the causal effect in observational study. Yet, it should also be stressed that the conclusions interpreted above are limited to the simulation's context. It cannot be generalized to all causal inference problems in observational data. For example, the super-learner using data-adaptive machine learning algorithms may not perform well with the data used in this thesis's simulation due to the nature of data (i.e., sample size, the dimensionality of the data). Besides, only binary treatment is examined throughout this thesis. This could be considered as one of the restriction of this thesis, since there are more type of treatments, for example, continuous treatments and multivariate treatments. Therefore, a comparative study in high-dimensional data, with different type of treatment mechanisms could be potential future research since TMLE's ultimate motivation is to be applied in such scenarios.

References

Allan, Victoria, Sreeram V Ramagopalan, Jack Mardekian, Aaron Jenkins, Xiaoyan Li, Xianying Pan, and Xuemei Luo. 2020. Propensity score matching and inverse probability of treatment weighting to address confounding by indication in comparative effectiveness research of oral anticoagulants. *Journal of Comparative Effectiveness Research*, 9(9):603–614.

Austin, Peter C. 2011. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behavioral Research*, 46(3):399–424.

Austin, Peter C., Jack V. Tu, and Douglas S. Lee. 2010. Logistic regression had superior performance compared with regression trees for predicting in-hospital mortality in patients hospitalized with heart failure. *Journal of Clinical Epidemiology*, 63(10):1145–1155.

Bang, Heejung and James M. Robins. 2005. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962–973.

Chernozhukov, Victor, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. 2016. Double/Debiased Machine Learning for Treatment and Causal Parameters.

D. W. Hosmer and S. Lemeshow. 2016. Sample Data: Birth Weight Risk | Wolfram Data Repository.

Elze, Markus C., John Gregson, Usman Baber, Elizabeth Williamson, Samantha Sartori, Roxana Mehran, Melissa Nichols, Gregg W. Stone, and Stuart J. Pocock. 2017. Comparison of Propensity Score Methods and Covariate Adjustment: Evaluation in 4 Cardiovascular Studies.

Faraoni, David and Simon Thomas Schaefer. 2016. Randomized controlled trials vs. observational studies: Why not just live together?

Ferri-García, Ramón and María Del Mar Rueda. 2020. Propensity score adjustment using machine learning classification algorithms to control selection bias in online surveys. *PLoS ONE*, 15(4).

Funk, Michele Jonsson, Daniel Westreich, Chris Wiesen, Til Stürmer, M. Alan Brookhart, and Marie Davidian. 2011. Doubly robust estimation of causal effects. *American Journal of Epidemiology*, 173(7):761–767.

Gelman, Andrew and Jennifer Hill. 2006. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press.

Han, Feng. 2018. Doubly robust estimation of the causal effects in the causal inference with missing outcome data. *Journal of Ambient Intelligence and Humanized Computing*, pages 1–9.

Ho, Daniel, Kosuke Imai, Gary King, and Elizabeth Stuart. 2007. Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference. *Political Analysis*, 15.

Holland, Paul W. 1986. Statistics and causal inference. *Journal of the American Statistical Association*, 81(396):945–960.

Kang, Joseph D.Y. and Joseph L. Schafer. 2007. Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science*, 22(4):523–539.

King, Gary and Richard Nielsen. 2019. Why Propensity Scores Should Not Be Used for Matching. *Political Analysis*, 27(4):435–454.

Künzel, Sören R., Jasjeet S. Sekhon, Peter J. Bickel, and Bin Yu. 2019. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences of the United States of America*, 116(10):4156–4165.

Laan, M. and S. Dudoit. 2003. Unified Cross-Validation Methodology For Selection Among Estimators and a General Cross-Validated Adaptive Epsilon-Net Estimator: Finite Sample Oracle Inequalities and Examples. *undefined*.

van der Laan, Mark J. and Richard J. C. M. Starmans. 2014. Entering the Era of Data Science: Targeted Learning and the Integration of Statistics and Computational Data Analysis. *Advances in Statistics*, 2014:1–19.

Lee, Brian K., Justin Lessler, and Elizabeth A. Stuart. 2010. Improving propensity score weighting using machine learning. *Statistics in Medicine*, 29(3):337–346.

Lendle, Samuel D., Bruce Fireman, and Mark J. Van Der Laan. 2013. Targeted maximum likelihood estimation in safety analysis. *Journal of Clinical Epidemiology*, 66(8 SUPPL.8):S91–S98.

Luque-Fernandez, Miguel Angel, Michael Schomaker, Bernard Rachet, and Mireille E. Schnitzer. 2018. Targeted maximum likelihood estimation for a binary treatment: A tutorial. *Statistics in Medicine*, 37(16):2530–2546.

Mark van der Laan. 2017. Targeted Learning: The Link From Statistics To Data Science. *STAtOR*, pages 12–16.
McConnell, K. John and Stephan Lindner. 2019. Estimating treatment effects with machine learning. *Health Services Research*, 54(6):1273–1282.
Naimi, Ashley I., Stephen R. Cole, and Edward H. Kennedy. 2017. An introduction to g methods. *International Journal of Epidemiology*, 46(2):756–762.
Polley, Eric and Mark van der Laan. 2010. Super Learner In Prediction. *U.C. Berkeley Division of Biostatistics Working Paper Series*.
Robins, James M., Andrea Rotnitzky, and Lue Ping Zhao. 1994. Estimation of Regression Coefficients When Some Regressors Are Not Always Observed. *Journal of the American Statistical Association*, 89(427):846.
Rosenbaum, Paul R. and Donald B. Rubin. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55.
Rubin, Donald B. 2005. Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 100(469):322–331.
Schuler, Megan S. and Sherri Rose. 2017. Targeted maximum likelihood estimation for causal inference in observational studies. *American Journal of Epidemiology*, 185(1):65–73.
Słoczyński, Tymon and Jeffrey M. Wooldridge. 2018. A General Double Robustness Result For Estimating Average Treatment Effects. *Econometric Theory*, 34(1):112–133.
Snowden, Jonathan M., Sherri Rose, and Kathleen M. Mortimer. 2011. Implementation of G-computation on a simulated data set: Demonstration of a causal inference technique. *American Journal of Epidemiology*, 173(7):731–738.
Van Der Laan, Mark J., Eric C. Polley, and Alan E. Hubbard. 2007. Super learner. *Statistical Applications in Genetics and Molecular Biology*, 6(1).
Van Der Laan, Mark J. and Daniel Rubin. 2006. Targeted Maximum Likelihood Learning. *U.C. Berkeley Division of Biostatistics Working Paper Series*.
Van Der Laan, Mark J. and Sherri Rose. 2011. *Targeted Learning - Causal Inference for Observational and Experimental Dataand Experimental Data | Mark J. van der Laan | Springer*. Springer-Verlag New York.
Wager, Stefan and Susan Athey. 2018. Estimation and Inference of Heterogeneous Treatment Effects using Random Forests. *Journal of the American Statistical Association*, 113(523):1228–1242.