Chapter 21
Propensity-Score-Based Estimators and C-TMLE

Jasjeet S. Sekhon, Susan Gruber, Kristin E. Porter, Mark J. van der Laan

In order to estimate the average treatment effect $E_0[E_0(Y \mid A = 1, W) - E_0(Y \mid A = 0, W)]$ of a single time-point treatment A based on observing n i.i.d. copies of $O = (W, A, Y)$, one might use inverse probability of treatment (i.e., propensity score) weighting of an estimator of the conditional mean of the outcome (i.e., response surface) as a function of the pretreatment covariates. Alternatively, one might use a TMLE defined by a choice of initial estimator, a parametric submodel that codes fluctuations of the initial estimator, and a loss function used to determine the amount of fluctuation, where either the choice of submodel or the loss function will involve inverse probability of treatment weighting. Asymptotically, such double robust estimators may have appealing properties. They can be constructed such that if either the model of the response surface or the model of the probability of treatment assignment is correct, the estimatosr will provide a consistent estimator of the average treatment effect. And if both models are correct, the weighted estimator will be asymptotically efficient. Such estimators are called double robust and locally efficient (Robins et al. 1994, 1995; Robins and Rotnitzky 1995; van der Laan and Robins 2003).

By factorization of the likelihood of O into a factor that identifies the average treatment effect, and the conditional probability of treatment, given the covariates W (i.e, the treatment mechanism), estimation of the propensity score $g_0(1 \mid W) = P_0(A = 1 \mid W)$ should be based solely on the log-likelihood of the treatment mechanism. In particular, estimation of g_0 should not involve examining the data on the final outcome Y. The double robust estimators exhibit a particularly interesting and useful type of asymptotics with respect to the choice of estimator of the propensity score. Due to this factorization of the likelihood, if one uses a maximum likelihood estimator of g_0 according to a particular model for g_0, then the influence curve of the double robust estimator equals the influence curve it would have had if g_0 were known and not estimated, minus a projection term whose size is implied by the size of the model for g_0 (van der Laan and Robins 2003). As a consequence of this result, the larger the model for g_0, the more efficient the double robust estimator will be. In addition, an estimator of the variance of the double robust estimator that ignores the

M.J. van der Laan and S. Rose, *Targeted Learning: Causal Inference for Observational and Experimental Data*, Springer Series in Statistics, DOI 10.1007/978-1-4419-9782-1_21,

fact that g_0 was estimated is asymptotically conservative, and thus provides valid conservative confidence intervals and tests. In the special case that the estimator of the response surface is consistent, it follows that the influence curve of the double robust estimator is not affected at all by the choice of estimator of g_0 (the above-mentioned projection term equals zero): under this assumption, even estimators that rely on sequential learning based on the log-likelihood of g_0, will not affect the statistical inference. Of course, such statements are not warranted if the choice of model for g_0 is based on examining the data on the final outcomes Y.

In finite samples, however, double robust estimators can increase variance and bias for the average treatment effect, relative to the estimator of the average treatment effect based on the unweighted estimator of the outcome, especially when some observations have an extreme probability of assignment to treatment, corresponding with practical or theoretical violations of the positivity assumption. Recall, that the positivity assumption states that the conditional probability of treatment assignment is bounded away from 0 and 1 for all covariate values. As a result, Kang and Schafer (2007) and Freedman and Berk (2008) warn against the routine use of estimators that rely on IPCW, including double robust estimators. This is in agreement with the past and ongoing literature defining and analyzing this issue (Robins 1986, 1987a, 2000; Robins and Wang 2000; van der Laan and Robins 2003), simulations demonstrating the extreme sparsity bias of IPCW estimators (e.g., Neugebauer and van der Laan 2005), diagnosing violations of the positivity assumptions in response to this concern (Kish 1992; Wang et al. 2006; Cole and Hernan 2008; Bembom and van der Laan 2008; Moore et al. 2009; Petersen et al. 2010), data-adaptive selection of the truncation constant to control the influence of weighting (Bembom and van der Laan 2008), and selecting parameters that rely on realistic assumptions (van der Laan and Petersen 2007a; Petersen et al. 2010).

One problem with reliance on the propensity score is that it may condition on variables that are either unrelated to the outcome of interest or only weakly related. Adding a pretreatment variable that is unrelated to the outcome but related to treatment (i.e., an instrument) to a propensity score model may increase bias. If the relationships between the variables are linear, bias will always be increased (Bhattacharya and Vogt 2007; Wooldridge 2009). In the nonparametric case, the direction of the bias is less straightforward, but increasing bias is a real possibility and expected (Pearl 2010a). This type of bias implied by such bias-amplifying variables has been termed Z-bias. In the nonparametric case, bias may result even when there are no unobserved confounders.

Including variables in a propensity score model that are unrelated to the outcomes of interest also exacerbates the problem of small or large estimated probabilities of treatment assignment. Such probabilities make inverse probability of treatment-weighted estimators unstable, and may, in finite samples, appear to cause violations of the positivity assumption. But these violations may be innocuous because the variable causing the violation may be unrelated to the outcome, and, therefore, one need not condition on it.

Even ignoring concerns about Z-bias, another consideration may lead one to want to examine the outcomes in order to decide on the fit of the propensity score: It is

often impossible to balance all of the theoretically plausible confounders in a given sample. In such cases, one cannot estimate the treatment effect without making functional form assumptions outside the support of the data. However, if the pretreatment variables that we cannot balance are unrelated to the outcomes, it may be possible to make progress. Of course, it would be preferable if a priori our scientific beliefs were sufficient to exclude the problematic pretreatment variables, but that is often unrealistic.

The foregoing paragraphs present us with a conundrum. For effective bias reduction based on the propensity score, we need to examine outcomes to include covariates in the propensity score fit that are meaningful confounders, but that contradicts the goal of fitting a propensity score, conditional on all the available covariates, and, as a consequence, it will alter the statistical understanding of the estimator, and its inference in a fundamental way. If we go the route of modeling the propensity score and the response surface together, then it is essential that the stability problems created by weighting be resolved. In addition, this will need to be done with an a priori specified algorithm.

C-TMLEs have desirable features that help to mitigate many of the concerns regarding the use of double robust estimators. First, the estimation of the response surface is automated by an a priori specified machine learning algorithm, such as the super learner. Second, the whole procedure for selecting the propensity score fit is automated. That is, the C-TMLE is an a priori specified estimator of the average treatment effect. Instead of giving the designer the freedom to select a fit of the propensity score based on the orthogonal log-likelihood of the propensity score before committing to an estimator of the average treatment effect that will use this fit of the propensity score, the C-TMLE lets an a priori specified machine determine this choice based on all the available data. It is important to note that choices that define the manner in which the C-TMLE fits the propensity score may still be based on inspection of the ancillary log-likelihood of the propensity score (i.e., not examining final outcome data).

Third, C-TMLE, by construction, only aims to include the variables in the propensity score that are related to the outcome of interest. More precisely, C-TMLE only includes variables in the propensity score if they are inadequately adjusted for by the fit of the response surface. Thus, concerns about Z-bias are reduced. Fourth, the estimated probabilities of treatment assignment are less likely to be extreme because typically fewer variables and fewer problematic variables will be included in the propensity score model by C-TMLE than by noncollaborative methods. The theory of collaborative double robustness provides the theoretical underpinning of the C-TMLE algorithm, showing that full bias reduction is achieved by using a propensity score that only adjusts for the covariates that explain the residual bias between the initial estimator of the response surface and the true response surface (Appendix A). Finally, as outlined in Chap. 7, both TMLE and C-TMLE can make use of a logistic fluctuation to make sure that the fit of the response surface either respects a priori known bounds of the continuous outcome or enforces the bounds implied by the range of the continuous outcome observed in the data. With a logistic fluctuation, the TMLE and C-TMLE will have the predicted outcome bounded to its

observed range even in finite samples. This bounding stabilizes the estimator because the influence function of the estimator is bounded, even in finite samples, and this controls the instability in the context of sparsity.

To summarize, Z-bias and the inability to balance all of the a priori plausible confounders suggests that one may want to use an a priori defined estimator that views fitting the propensity score as a task that needs to be carried out in collaboration with fitting the response surface. C-TMLEs have a number of desirable advantages relative to other double robust estimators that view fitting the propensity score as an external task based on the orthogonal log-likelihood of the propensity score that ignores the outcome data.

Chapter Summary

Above, we made a compelling case for the C-TMLE. Statistical properties of estimators, and thereby the comparison between different estimation procedures, are not affected by the choice of causal model, only by the statistical target parameter and the statistical model. Nonetheless, causal models represent an intrinsic component of our road map for targeted learning of causal effects. Chapter 2 was devoted to the SCM, and this causal model was repeatedly invoked in this book. In this causal model, counterfactuals, and thereby the causal quantities of interest, were derived from the SCM. An important and popular alternative causal model directly states the existence of counterfactuals of scientific interest as the main assumption (often defined in words in terms of an experiment), which avoids the representation of assumptions in terms of structural equations. This is called the Neyman–Rubin model and is presented in the next section. The debate over which estimator to select for estimation of a causal effect should not be concerned with the choice of causal model.

We then reexamine prominent Monte Carlo simulations by implementing the C-TMLE and TMLE. These estimators may overcome the more detailed, but in some sense second-order, concerns regarding the instability of double robust estimators under sparseness. Specifically, we reanalyze and extend the simulations of Kang and Schafer (2007), which were designed to highlight the limitations of double robust estimators. The outcome is a linear function of the covariates, and the error term is small relative to the size of the coefficients. In addition, the missingness mechanism results in many positivity violations in finite samples. For the double robust estimators Kang and Schafer considered, the double robust weighted least squares and A-IPTW result in volatile estimates due to the high-leverage data points generated by large weights. We also examine the simulations of Freedman and Berk (2008). These simulations were originally designed to demonstrate how weighting by the propensity score can result in highly unstable estimates in conditions that are less extreme than those of Kang and Schafer. In both simulation studies, we show that the TMLEs, and, in particular, the C-TMLEs, perform well.

21.1 Neyman–Rubin Causal Model and Potential Outcomes

The Neyman–Rubin causal model consists of more than simply the notation for potential outcomes that was originated by Neyman (1923). Rubin and others, such as Cochran, Holland, and Rosenbaum, have developed a general framework that helps to clarify some important issues of inference and design. Beginning with Rubin (1974), the model began to unify how one thinks about observational and experimental studies, and it gives a central place to the treatment assignment mechanism.

The Neyman–Rubin framework has become increasingly popular in many fields, including statistics (Holland 1986; Rubin 1974, 2006; Rosenbaum 2002), medicine (Rubin 1997; Christakis and Iwashyna 2003), economics (Dehejia and Wahba 1999, 2002; Galiani et al. 2005; Abadie and Imbens 2006), political science (Herron and Wand 2007; Imai 2005; Sekhon 2008b), sociology (Morgan and Harding 2006; Diprete and Engelhardt 2004; Winship and Morgan 1999; Smith 1997), and even law (Rubin 2002).The framework originated with Neyman's model, which is nonparametric for a finite number of treatments. In the case of one treatment and one control condition, each unit has two potential outcomes, one if the unit is treated and the other if untreated. A causal effect is defined as the difference between the two potential outcomes, but only one of the two potential outcomes is observed. Rubin (1974, 2006) developed the model into a general framework for causal inference with implications for observational research. Holland (1986) wrote an influential review article that highlighted some of the philosophical implications of the framework. Consequently, instead of the "Neyman–Rubin model," the model is often simply called the Rubin causal model (e.g., Holland 1986) or sometimes the Neyman–Rubin–Holland model (e.g., Brady 2008) or the Neyman–Holland–Rubin model (e.g., Freedman 2006).

The intellectual history of the Neyman–Rubin model is the subject of some controversy (e.g., Freedman 2006; Rubin 1990; Speed 1990). Neyman's 1923 article never mentions the random assignment of treatments. Instead, the original motivation was an urn model, and the explicit suggestion to use the urn model to physically assign treatments is absent from the paper (Speed 1990). It was left to R. A. Fisher in the 1920s and 1930s to note the importance of the physical act of randomization in experiments. Fisher first did this in the context of experimental design in his 1925 book, expanded on the issue in a 1926 article for agricultural researchers, and developed it more fully and for a broader audience in his 1935 book *The Design of Experiments*.

This gap between Neyman and Fisher points to the fact that there was something absent from the Neyman mathematical formulation, which was added later, even though the symbolic formulation was complete in 1923. What those symbols *meant* changed. And in these changes lies what is causal about the Neyman–Rubin model—i.e., a focus on the mechanism by which treatment is assigned. The Neyman–Rubin model is more than just the math of the original Neyman model. Obviously, it relies not on an urn model motivation for the observed potential outcomes but, for experiments, a motivation based on the random assignment of treatment. And for observational studies, one relies on the assumption that the assign-

ment of treatment can be treated as if it were random. In either case, the mechanism by which treatment is assigned is of central importance. And the realization that the primacy of the assignment mechanism holds true for observational data no less than for experimental, is due to Rubin (1974).

The basic setup of the Neyman model is simple. Let A_i be a treatment indicator: 1 when i is in the treatment regime and 0 otherwise. But an additional assumption must be made to link potential outcomes to observed outcomes. The most common assumption used is the one of "no interference between units" (Cox 1958, Sect. 2.4). With this assumption, one can assume that the observed outcome for observation i is

$$Y_i = A_i Y_{1,i} + (1 - A_i) Y_{0,i}, \tag{21.1}$$

where $Y_{1,i}$ denotes the potential outcome for unit i if the unit receives treatment and $Y_{0,i}$ denotes the potential outcome for unit i in the control regime. The treatment effect for observation i is defined by $\tau_i = Y_{1,i} - Y_{0,i}$. Causal inference is a missing-data problem because $Y_{1,i}$ and $Y_{0,i}$ are never both observed.

The "no interference between units" is often called the stable unit treatment value assumption (SUTVA). SUTVA implies that the potential outcomes for a given unit do not vary with the treatments assigned to any other unit, and that there are not different versions of treatment (Rubin 1978). The mapping from potential outcomes to observed outcomes is not a primitive in the potential-outcomes framework. This point is often missed. Equation (21.1) only follows because of the no interference assumption. Otherwise, Y_i may depend on A_j, $Y_{0,j}$, and $Y_{1,j}$, where $j \neq i$. Therefore, one may argue that there is a structural model embedded in the Neyman model as commonly used, although a simple one with clear behavioral implications.

The next key part of the Neyman–Rubin model is the assignment mechanism. It is the process by which the potential outcomes are missing. The treatment assignment mechanism may satisfy the no unmeasured confounding assumption:

$$P(A \mid W, Y_0, Y_1) = P(A \mid W), \tag{21.2}$$

where W are some confounders. If the randomization is not conditional on W, then $P(A \mid W, Y_0, Y_1) = P(A \mid W) = P(A)$. Beyond randomization, one wishes that

$$0 < P(A_i = 1 \mid W_i) < 1. \tag{21.3}$$

Equation (21.3) is also referred to as the positivity assumption, as discussed in detail in Chaps. 2 and 10. It is a common support condition.

Equations (21.2) and (21.3) make clear that randomized experiments are free of any dependence between the treatment *and* potential outcomes, conditional on observables that were used to define the randomization probabilities. Before Rubin (1975), the potential-outcomes framework was used by various authors to formalize randomized experiments, but never observational studies. Freedman (2006), Rubin (2008), and Sekhon (2010) review some of the relevant history.

It was unprecedented when Rosenbaum and Rubin (1983) used Eqs. (21.2) (ignorability) and (21.3) (common support) to define "strong ignorability." Ignorability

dates from Rubin (1976). Importantly, these concepts were applied, for the first time, to observational data and not just experimental data. This gave a formal language to the tradition of thinking of observational studies as broken experiments, where many of the lessons of experimental design are maintained. This tradition goes back to at least Cochran (1965) – also see Cochran and Rubin (1973) and Cochran (1983). Rubin and his various co-authors formalized and extended it to become a way of thinking about causality. There are no observed outcomes in Eqs. (21.2) and (21.3), but they define an assumption by which the average treatment effect may be identified: selection on observables.

If the principle of defining an estimator in terms of an a priori fitted propensity score, in which the fitting process can be flexible but is ignorant of the outcomes, is a valid one, then the statistical properties of double robust estimators (such as the TMLE with an externally fitted treatment mechanism) that live by this principle should be fully competitive with a (double robust) C-TMLE that ignores this principle and lives by the principle of targeted learning that views the goal of the fit of the propensity score as an ingredient to obtain a best estimator of the target quantity.

Three questions arise that we wish to address. First, can an automated estimation strategy be used effectively? Second, can the TMLE and the C-TMLE overcome the demonstrated instability problems of other double robust estimators? Third, overall, which one wins, the TMLE, which lives by the principle of using an externally fitted propensity score, or the C-TMLE, which fits the propensity score as an ingredient for targeted fitting of the response surface?

21.2 Kang and Schafer Censored-Data Simulations

In this section, motivated by Kang and Schafer (2007) (hereafter, KS) and a response by Robins et al. (2007b), we compare the performance of TMLEs to that of estimating-equation-based double robust estimators, in the context of sparsity. This set of simulations was originally designed by KS to highlight the stability problems of double robust estimators. We will demonstrate that TMLEs can perform well in these simulations. The KS simulations focus on the problem of estimating a population mean from censored data. The data are CAR but not completely at random. We explore the relative performance of the estimators under the original KS simulation and a number of alternative data-generating distributions that involve different and stronger types of violations of the positivity assumption. These new simulation settings were designed to provide more diverse and challenging test cases for evaluating robustness and thereby finite sample performance of the different estimators.

Original Kang and Schafer simulation. KS considered n i.i.d. units of $O = (W, \Delta, \Delta Y) \sim P_0$, where W is a vector of four baseline covariates and Δ is an indicator of whether the continuous outcome, Y, is observed. KS were interested in estimating $\mu(P_0) = E_0(Y) = E_0[E_0(Y \mid \Delta = 1, W)]$. Let $(Z_1, \ldots, Z_4)$ be independent normally distributed random variables with mean zero and variance one.

The covariates W we actually observe were generated as follows: $W_1 = \exp(Z_1/2)$, $W_2 = Z_2/(1 + \exp(Z_1)) + 10$, $W_3 = (Z_1 Z_3/25 + 0.6)^3$, and $W_4 = (Z_2 + Z_4 + 20)^2$. The outcome Y was generated as $Y = 210 + 27.4Z_1 + 13.7Z_2 + 13.7Z_3 + 13.7Z_4 + N(0, 1)$. From this one can determine that the conditional mean $\bar{Q}_0$ of Y, given W, which equals the same linear regression in $Z_1(W), \ldots, Z_4(W)$, where $Z_j(W)$, $j = 1, \ldots, 4$, are the unique solutions of the four equations above in terms of $W = (W_1, \ldots, W_4)$. Formally, $\bar{Q}_0(W) = E_0[E_0(Y \mid Z) \mid W]$. Thus, if the data analyst had been provided the functions $Z_j(W)$, then the true regression function was linear in these functions, but the data analyst is measuring the terms W_j instead.

The other complication of the data-generating distribution is that Y is subject to missingness, and the true censoring mechanism, denoted by $g_0(1 \mid W) = P_0(\Delta = 1 \mid W)$, is given by $g_0(1 \mid W) = \text{expit}(-Z_1(W) + 0.5Z_2(W) - 0.25Z_3(W) - 0.1Z_4(W))$. With this data-generating mechanism, the average response rate is 0.50. Also, the true population mean is 210, while the mean among respondents is 200. These values indicate a small selection bias.

In these simulations, a linear main term model in the main terms $(W_1, \ldots, W_4)$ for either the outcome-regression or missingness mechanism is misspecified, while a linear main term model in the main terms $(Z_1(W), \ldots, Z_4(W))$ would be correctly specified. Note that there are finite sample violations of the positivity assumption given in Eq. (21.3). Specifically, we find $g_0(\Delta = 1 \mid W) \in [0.01, 0.98]$, and the estimated missingness probabilities $g_n(\Delta = 1 \mid W)$ were observed to fall in the range $[4 \times 10^{-6}, 0.97]$.

Modified Kang and Schafer simulation 1. In the KS simulation, when $\bar{Q}_0$ or g_0 is misspecified, the misspecifications are small. The selection bias is also small. Therefore, we modified the KS simulation in order to increase the degree of misspecification and to increase the selection bias. This creates a greater challenge for estimators and better highlights their relative performance. As before, let Z_j be i.i.d. $N(0, 1)$. The outcome Y was generated as $Y = 210 + 50Z_1 + 25Z_2 + 25Z_3 + 25Z_4 + N(0, 1)$. The covariates actually observed by the data analyst are now given by the following functions of $(Z_1, \ldots, Z_4)$: $W_1 = \exp(Z_1^2/2)$, $W_2 = 0.5Z_2/(1 + \exp(Z_1^2)) + 3$, $W_3 = (Z_1^2 Z_3/25 + 0.6)^3 + 2$, and $W_4 = (Z_2 + 0.6Z_4)^2 + 2$. From this, one can determine the true regression function $\bar{Q}_0(W) = E_0(E_0(Y \mid Z) \mid W)$. The missingness indicator was generated as follows: $g_0(1 \mid W) = \text{expit}(-2Z_1 + Z_2 - 0.5Z_3 - 0.2Z_4)$. A misspecified fit is now obtained by fitting a linear or logistic main term regression in $W_1, \ldots, W_4$, while a correct fit is obtained by providing the user with the terms $Z_1, \ldots, Z_4$, and fitting a linear or logistic main term regression in $Z_1, \ldots, Z_4$. With these modifications, the population mean is again 210, but the mean among respondents is 184.4. With these modifications, we have a higher degree of practical violation of the positivity assumption: $g_0(\Delta = 1 \mid W) \in [1.1 \times 10^{-5}, 0.99]$ while the estimated probabilities, $g_n(\Delta = 1 \mid W)$, were observed to fall in the range $[2.2 \times 10^{-16}, 0.87]$.

Modified Kang and Schafer simulation 2. Here we made one additional change to the modified simulation 1: We set the coefficient in front of Z_4 in the true regression

of Y on Z equal to zero. Therefore, while Z_4 is still associated with missingness, it is not associated with the outcome, and is thus not a confounder. Given $(W_1, \ldots, W_3)$, W_4 is not associated with the outcome either, and therefore as misspecified regression model of $\bar{Q}_0$ we use a main term linear regression in (W_1, W_2, W_3). This modification to the KS simulation enables us to take the debate on the relative performance of double robust estimators one step further, by addressing a second key challenge of the estimators: They often include nonconfounders in the censoring mechanism estimator. This unnecessary inclusion could unnecessarily introduce positivity violations. Moreover, this can itself introduce substantial bias and inflated variance, sometimes referred to as Z-bias. While this problem is not presented in the Kang and Schafer paper and responses, it is highlighted in the literature, including Bhattacharya and Vogt (2007), Wooldridge (2009), and Pearl (2010a). As discussed earlier, the C-TMLE provides an innovative approach for estimating the censoring mechanism, preferring covariates that are associated with the outcome and censoring, without "data snooping."

21.2.1 Estimators

As illustrated in KS and Robins et al. (2007b), semiparametric efficient double robust estimators typically rely on IPCW. These weights will be very large when there are violations of the positivity assumption. As a benchmark, KS compare all estimators in their article to the ordinary least squares estimator:

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

where $\bar{Q}_n$ is the least squares estimator of $\bar{Q}_0$ according to a main term linear regression model $m_\beta(W) = \beta W^\top$, only using the observations with $\Delta_i = 1$.

Kang and Schafer present comparisons of several double robust (and nondouble robust) estimators. We focus on the weighted least squares (WLS) estimator and the A-IPCW estimator. The WLS estimator is defined as

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

where m_β is a linear regression model for $\bar{Q}_0$ and β_n is an IPCW linear regression estimator given by

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

The A-IPCW estimator is defined as

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

We compare these estimators with the TMLE and the C-TMLE with logistic fluctuation for a continuous outcome (Chap. 7). These estimators are guaranteed to stay within the global bounds of the model, which is essential when $g_0(1 \mid W)$ has values close to 0. The logistic fluctuation TMLE for continuous $Y \in (a, b)$ involves defining a normalized outcome $Y^* = (Y-a)/(b-a) \in (0, 1)$, computing the TMLE of $E_0(Y^*)$, and transforming back this TMLE into a TMLE of $E_0(Y) = (b-a)E_0(Y^*) + a$. The TMLE with logistic fluctuation requires setting a range $[a, b]$ for the outcomes Y. If such knowledge is available, one simply uses the known values. If Y is not subject to missingness, then one would use the minimum and maximum of the empirical sample, which represents a very accurate estimator of the range. In these simulations, Y is subject to informative missingness such that the minimum or maximum of the biased sample represents a biased estimate of the range, resulting in a small unnecessary bias in the TMLE (negligible relative to MSE). We enlarged the range of the complete observation by a factor of 1.1, which seemed to remove most of the unnecessary bias. We expect that some improvements can be obtained by incorporating a valid estimator of the range that takes into account the informative missingness, but such second order improvements are outside the scope of this chapter. The TMLE of $E_0(Y^*)$ involves obtaining an initial estimator of $E_0(Y^* \mid W, \Delta = 1)$, representing it as a logistic function, and subsequently fluctuating it according to the logistic fluctuation function. Let $\bar{Q}_n^0 \in (0, 1)$ be an initial estimator of $E_0(Y^* \mid \Delta = 1, W)$ obtained by regressing Y^* onto W among the observations with $\Delta_i = 1$. Consider the logistic fluctuation working model

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

where $H_{g,n}^*(1, W) = 1/g_n(1 \mid W)$. One estimates the amount of fluctuation ϵ with maximum likelihood estimation using logistic regression software for binary outcomes. One can use standard software for this fluctuation, ignoring that Y^* is not binary. (See Chap. 7 for the proof that the binary outcome log-likelihood loss function is the correct loss function in this case.) This now defines the first-step TMLE $\bar{Q}_n^1 = \bar{Q}_n^0(\epsilon_n)$, which is also the TMLE of $\bar{Q}_0$. The TMLE of $E_0(Y^*)$ is now given by the corresponding substitution estimator $\mu_{TMLE,n}^* = \frac{1}{n}\sum_i \bar{Q}_n^*(W_i)$. The latter estimator maps into the desired TMLE $\mu_{TMLE,n} = (b-a)\mu_{TMLE,n}^* + a$ of $\mu_0 = E_0(Y)$.

The C-TMLE $\mu_{C-TMLE,n}$ is defined in Chap. 20 for the mean outcome $E_0 Y_1$ based on $O = (W, A, Y)$, but now treating $A = \Delta$. The C-TMLE differs from the standard TMLE above in its estimation procedures for $\bar{Q}_0$ and g_0. The TMLE fluctuates an initial estimator of $\bar{Q}_0$ using an external estimate of g_0, while the C-TMLE estimate considers a sequence of subsequent TMLE updates of this initial estimator indexed by increasingly nonparametric estimators of g_0 and, based on the "log-likelihood" for $\bar{Q}_0$ of these candidate TMLEs, data-adaptively determines the desired TMLE and, in particular, the desired fit of g_0. The C-TMLE involves building an estimator of the distribution of Δ as a function of a set of covariates that are still predictive

of Y, after taking into account the initial estimator. That is, the C-TMLE relies on a collaboratively estimated g_n that at times only targets a true conditional distribution of Δ, given a *reduction* of W, yet delivers full bias reduction.

21.2.2 Results

For the three simulations described above, the OLS, WLS, A-IPCW, TMLE, and C-TMLE were used to estimate $\mu(P_0)$ from 250 samples of size 1000. We evaluated the performance of the estimators by their bias, variance, and MSE. We compared the estimators of $\mu(P_0)$ using different specifications of the estimators of $\bar{Q}_0$ and g_0. In the tables below, *CC* indicates that the estimators of both were specified correctly; *CM* indicates that the estimator of $\bar{Q}_0$ was correct, but the estimator for g_0 was misspecified; *MC* indicates that the estimator for $\bar{Q}_0$ was misspecified, but the estimator for g_0 was correct; and *MM* indicates both estimators were misspecified.

For all estimators, we compared results with $g_n(1 \mid W) \in [0, 1]$ by also truncating $g_n(1 \mid W)$ from below at three different levels: 0.010, 0.025, and 0.050. We note that neither KS nor Robins et al. (2007b) included bounding $g_n(1 \mid W)$ from below when applying their estimators. In any given application, it is difficult to determine which bounds to use, but the theory teaches us that double robust estimators can only be consistent if $g_n(1 \mid W)$ stays bounded away from zero, even if the true g_0 is

Table 21.1 Simulation results, 250 samples of size 1000. (a) Kang and Schafer simulation. (b) Modified simulation 1. (c) Modified simulation 2

	CC			*CM*			*MC*			*MM*		
(a)	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE
OLS	−0.09	1.4	1.4	−0.09	1.4	1.4	−0.93	2.0	2.8	−0.93	2.0	2.8
WLS	−0.09	1.4	1.4	−0.09	1.4	1.4	0.10	1.8	1.8	−3.0	2.1	11
A-IPCW	−0.09	1.4	1.4	−0.10	1.4	1.5	0.04	2.5	2.5	−8.8	2e+2	3e+2
TMLE	−0.10	1.4	1.4	−0.11	1.4	1.4	−0.09	2.1	2.1	−4.6	3.6	25
C-TMLE	−0.10	1.4	1.4	−0.11	1.4	1.4	0.09	1.8	1.8	−1.5	2.8	5.0
(b)												
OLS	−0.17	4.7	4.7	−0.17	4.7	4.7	−36	17	1e+3	−36	17	1e+3
WLS	−0.16	4.7	4.7	−0.16	4.7	4.7	−4.4	42	61	−35	16	1e+3
A-IPCW	−0.16	4.7	4.8	−0.16	4.7	4.7	−1.8	2e+2	2e+3	−35	17	1e+3
TMLE	−0.22	4.7	4.7	−0.23	4.7	4.7	−0.04	89	89	−34	6.5	1e+3
C-TMLE	−0.26	4.7	4.7	−0.22	4.7	4.7	−0.64	16	16	−34	6.6	1e+3
(c)												
OLS	−0.06	3.9	3.9	−0.06	3.9	3.9	−34	15	1e+3	−34	15	1e+3
WLS	−0.06	4.0	3.9	−0.06	3.9	3.9	−3.6	40	53	−33	15	1e+3
A-IPCW	−0.05	4.0	4.0	−0.06	3.9	3.9	−1.1	2e+2	2e+3	−33	16	1e+3
TMLE	−0.10	3.9	3.9	−0.11	3.9	3.9	0.15	76	76	−32	5.6	1e+3
C-TMLE	−0.14	3.9	3.9	−0.11	3.9	3.9	−0.88	11	11	−33	5.8	1e+3

Table 21.2 Kang and Schafer simulation results by truncation level of g_n, 250 samples of size 1000

Bound on g_n	*CC*			*CM*			*MC*			*MM*		
by estimator	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE
OLS	−0.09	1.4	1.4	−0.09	1.4	1.4	−0.93	2.0	2.8	−0.9	2.0	2.8
WLS												
None	−0.09	1.4	1.4	−0.09	1.4	1.4	0.10	1.8	1.8	−3.0	2.1	11
0.010	−0.09	1.4	1.4	−0.09	1.4	1.4	0.10	1.8	1.8	−3.0	2.0	11
0.025	−0.09	1.4	1.4	−0.09	1.4	1.4	0.10	1.8	1.8	−2.9	2.0	11
0.050	−0.09	1.4	1.4	−0.09	1.4	1.4	0.11	1.8	1.8	−2.7	1.9	9.4
A-IPCW												
None	−0.09	1.4	1.4	−0.10	1.4	1.5	0.04	2.5	2.5	−8.8	2e+2	3e+2
0.010	−0.09	1.4	1.4	−0.09	1.4	1.4	0.04	2.5	2.5	−6.1	18	56
0.025	−0.09	1.4	1.4	−0.09	1.4	1.4	0.04	2.4	2.4	−4.9	6.1	30
0.050	−0.09	1.4	1.4	−0.09	1.4	1.4	0.08	2.3	2.3	−3.8	3.2	18
TMLE												
None	−0.10	1.4	1.4	−0.11	1.4	1.4	−0.09	2.1	2.1	−4.6	3.6	25
0.010	−0.10	1.4	1.4	−0.10	1.4	1.4	−0.09	2.1	2.1	−4.4	4.2	24
0.025	−0.10	1.4	1.4	−0.10	1.4	1.4	−0.09	2.1	2.1	−4.1	3.1	20
0.050	−0.10	1.4	1.4	−0.10	1.4	1.4	−0.06	2.0	2.0	−3.6	2.4	15
C-TMLE												
None	−0.10	1.4	1.4	−0.11	1.4	1.4	0.09	1.8	1.8	−1.5	2.8	5.0
0.010	−0.10	1.4	1.4	−0.10	1.4	1.4	0.09	1.7	1.7	−1.3	2.2	4.0
0.025	−0.10	1.4	1.4	−0.10	1.4	1.4	0.11	1.7	1.7	−1.4	2.3	4.2
0.050	−0.10	1.4	1.4	−0.10	1.4	1.4	0.10	1.8	1.8	−1.3	2.1	3.8

not bounded away from zero (e.g., van der Laan and Robins 2003). Therefore, only presenting the (interesting) results for not bounding at all (but $g_n(1 \mid W) \in [0, 1]$) provides insight about an estimator that should never be used in practice. Ideally, the choice of bounding $g_n(1 \mid W)$ should depend on, among other things, the data-generating process and the sample size, so that one desires an estimator adaptively determines the truncation level (such as particular implementations of C-TMLE).

Table 21.1 presents the simulation results without any bounding of g_n.The tables show that in all three simulations, the TMLE and C-TMLE with a logistic fluctuation achieve comparable or better MSE than the other estimators. When $\bar{Q}_n$ is misspecified, TMLE performs well and C-TMLE stands out with a much lower MSE. Together, the results from modified simulation 1 and modified simulation 2 show that the C-TMLEs have similar or superior performance relative to estimating-equation-based double robust estimators when not all covariates are associated with Y. At the same time, even in cases in which *all* covariates are associated with Y, the C-TMLE still performs well. Tables 21.2 and 21.3 compare results for each estimator when bounding g_n at different levels. We see that bounding g_n can improve the bias and variability of the estimators, often substantially. However, we also see that bounding can easily increase bias. The effect of bounding and the desired level

Table 21.3 Simulation results by truncation level of g_n, 250 samples of size 1000. (a) Modified Kang and Schafer simulation 1. (b) Modified Kang and Schafer simulation 2

	CC			*CM*			*MC*			*MM*		
(a)	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE
OLS	−0.17	4.7	4.7	−0.17	4.7	4.7	−36	17	1e+3	−36	17	1e+3
WLS												
None	−0.16	4.7	4.7	−0.16	4.7	4.7	−4.4	42	61	−35	16	1e+3
0.010	−0.16	4.7	4.7	−0.16	4.7	4.7	−4.6	39	60	−35	16	1e+3
0.025	−0.17	4.7	4.7	−0.16	4.7	4.7	−5.5	32	62	−35	16	1e+3
0.050	−0.17	4.7	4.7	−0.16	4.7	4.7	−7.3	25	78	−35	16	1e+3
A-IPCW												
None	−0.16	4.7	4.8	−0.16	4.7	4.7	−1.8	2e+2	2e+3	−35	17	1e+3
0.010	−0.16	4.7	4.7	−0.16	4.7	4.7	−3.7	74	88	−35	17	1e+3
0.025	−0.17	4.7	4.7	−0.16	4.7	4.7	−5.9	43	77	−35	17	1e+3
0.050	−0.17	4.7	4.7	−0.16	4.7	4.7	−8.8	28	1e+2	−35	17	1e+3
TMLE												
None	−0.22	4.7	4.7	−0.23	4.7	4.7	−0.04	89	89	−34	6.5	1e+3
0.010	−0.22	4.7	4.7	−0.22	4.7	4.7	0.71	53	54	−34	6.5	1e+3
0.025	−0.22	4.7	4.7	−0.22	4.7	4.7	1.0	22	23	−34	6.5	1e+3
0.050	−0.22	4.7	4.7	−0.22	4.7	4.7	−0.49	11	11	−34	6.5	1e+3
C-TMLE												
None	−0.26	4.7	4.7	−0.22	4.7	4.7	−0.64	16	16	−34	6.7	1e+3
0.010	−0.24	4.7	4.8	−0.22	4.7	4.7	−0.84	22	22	−34	6.7	1e+3
0.025	−0.24	4.7	4.7	−0.22	4.7	4.7	−1.5	12	14	−34	6.8	1e+3
0.050	−0.23	4.7	4.7	−0.22	4.7	4.7	−2.6	8.7	15	−34	6.8	1e+3
(b)												
OLS	−0.06	3.9	3.9	−0.06	3.9	3.9	−34	15	1e+3	−34	15	1e+3
WLS												
None	−0.06	4.0	3.9	−0.06	3.9	3.9	−3.6	40	53	−33	15	1e+3
0.010	−0.06	4.0	3.9	−0.06	3.9	3.9	−4.0	35	51	−33	15	1e+3
0.025	−0.06	4.0	3.9	−0.06	3.9	3.9	−4.9	29	53	−33	15	1e+3
0.050	−0.06	4.0	3.9	−0.06	3.9	3.9	−6.7	23	68	−33	15	1e+3
A-IPCW												
None	−0.05	4.0	4.0	−0.06	3.9	3.9	−1.1	2e+2	2e+2	−33	16	1e+3
0.010	−0.06	4.0	4.0	−0.06	3.9	3.9	−3.1	70	80	−33	16	1.e+3
0.025	−0.06	4.0	3.9	−0.06	3.9	3.9	−5.4	39	68	−33	16	1e+3
0.050	−0.06	3.9	3.9	−0.06	3.9	3.9	−8.3	26	94	−33	16	1e+3
TMLE												
None	−0.10	3.9	3.9	−0.11	3.9	3.9	0.15	76	76	−32	5.6	1e+3
0.010	−0.10	3.9	3.9	−0.10	3.9	3.9	0.95	43	44	−32	5.6	1e+3
0.025	−0.10	3.9	3.9	−0.10	3.9	3.9	1.3	18	19	−32	5.6	1e+3
0.050	−0.10	3.9	3.9	−0.10	3.9	3.9	−0.20	8.5	8.5	−32	5.6	1e+3
C-TMLE												
None	−0.14	3.9	3.9	−0.11	3.9	3.9	−0.88	11	11	−33	5.8	1e+3
0.010	−0.13	3.9	3.9	−0.10	3.9	3.9	−0.91	12	12	−33	6.1	1e+3
0.025	−0.12	3.9	3.9	−0.10	3.9	3.9	−1.4	8.5	10	−33	6.1	1e+3
0.050	−0.12	3.9	3.9	−0.10	3.9	3.9	−2.5	6.4	12	−33	6.0	1e+3

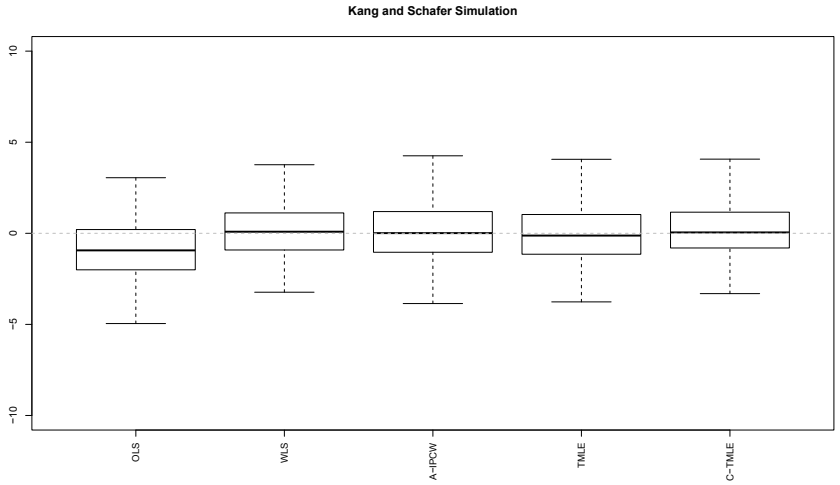

Fig. 21.1 Kang and Schafer simulation, MC, truncation level 0.025

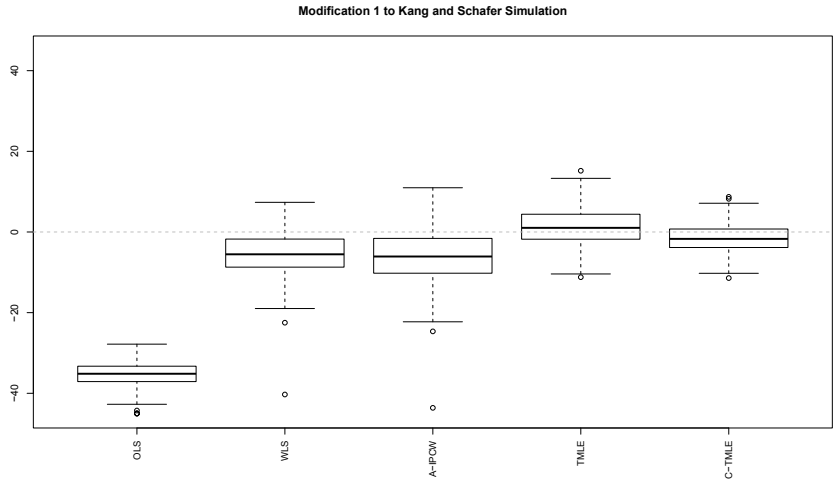

Fig. 21.2 Modified Kang and Schafer simulation 1, MC, truncation level 0.025

of bounding varies by estimator. It is important to note C-TMLE and TMLE are always well behaved. In no simulation do they show marked instability. C-TMLE performs particularly well. Results from the KS simulation, modified simulation 1, and modified simulation 2 are presented visually for MC with $g_n(1 \mid W)$ truncated from below at 0.025 in Figs. 21.1–21.3.

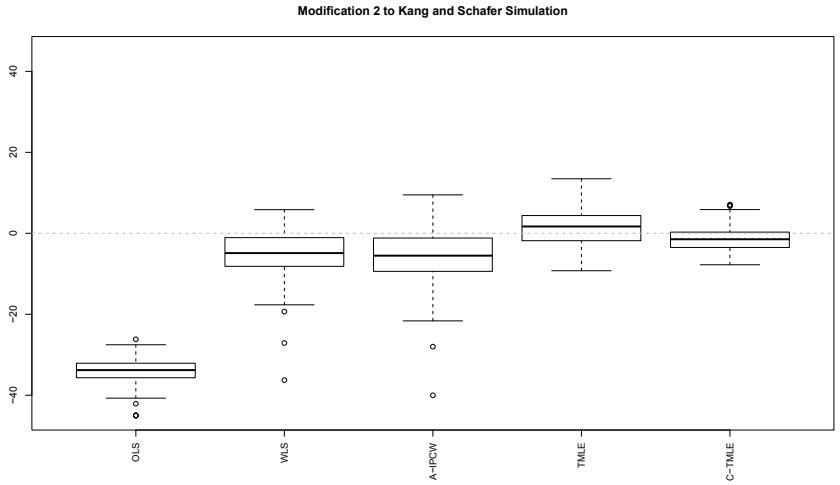

Fig. 21.3 Modified Kang and Schafer simulation 2, MC, truncation level 0.025

21.2.3 Super Learning and the Kang and Schafer Simulations

The misspecified formulation in KS can illustrate the benefits of coupling data-adaptive (super) learning with the TMLE. Results for the case that both $\bar{Q}_0$ and g_0 are inconsistently estimated indicate that the C-TMLE, constrained to use a main term regression model with misspecified covariates (W_1, W_2, W_3, W_4), has smaller variance than $\hat{\mu}_{OLS}$ but is more biased. The MSE of the TMLE is larger than the MSE of C-TMLE, with increased bias and variance. How would the estimation process be affected if we chose to act, based on the widespread understanding that models are seldom correctly specified and main term regressions generally fail to adequately capture the true relationships between predictors and an outcome, by turning to data-adaptive machine learning?

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

Table 21.4 Super learning simulation results, MM, $g_n(1 \mid W)$ truncated at 0.025

	Bias	Var	MSE
TMLE+ SL	−0.771	1.51	2.10
C-TMLE+ SL	−1.047	1.54	2.64

An important aspect of super learning is to ensure that the collection of prediction algorithms includes a variety of approaches for fitting the true function $\bar{Q}_0$ and g_0. For example, it is sensible to include a main terms regression algorithm in the super learner library. Should that algorithm happen to be correct, the super learner will behave as the main terms regression algorithm. It is also recommended to include algorithms that search over a space of higher-order polynomials, nonlinear models, and, for example, cubic splines. For binary outcome regression, as required for fitting g_0, classification algorithms such as classification and regression trees, support vector machines (Cortes and Vapnik 1995), and k-nearest-neighbor algorithms (Friedman 1994) could be added to the collection of algorithms. Super learning relies on the oracle property of V-fold cross-validation to asymptotically select the optimal convex combination of estimates obtained (Chap. 3).

Consider the misspecified scenario proposed by KS. The truth for both the outcome regression and the propensity for missingness regression is captured by a main terms linear regression of the outcome on Z_1, Z_2, Z_3, Z_4. This simple truth is virtually impossible to discover through the usual model selection approaches when the observed data consist of misspecified covariates $O = (W_1, W_2, W_3, W_4, \Delta, \Delta Y)$, given that $Z_1 = 2\log(W_1)$, $Z_2 = (W_2 - 10)(1 + 2W_1)$, $Z_3 = (25 \times (W_3 - 0.6))/(2\log(W_1))$, and $Z_4 = \sqrt[3]{W_4} - 20 - (W_2 - 10)(1 + 2W_1)$. This complexity illustrates the importance of including prediction algorithms that attack the estimation problem from a variety of directions. The collection of algorithms employed included glm, step, ipredbagg, DSA, earth, loess, nnet, svm, and k-nearest-neighbors. (We note that k-nearest-neighbors is only for binary outcomes, and it was used to estimate g only.)

In Table 21.4 we report the results for TMLE and C-TMLE based on 250 samples of size 1000, with predicted values for $g_n(1 \mid W)$ truncated from below at 0.025. The MSE for both estimators is smaller than the MSE of $\hat{\mu}_{OLS}$. The C-TMLE bias is slightly higher than the $\hat{\mu}_{OLS}$ bias, and TMLE is slightly better with respect to both bias and variance. More importantly, data-adaptive estimation improved efficiency of TMLE by a factor of 8.5. C-TMLE efficiency improved by a factor of 1.5.

21.3 Freedman and Berk Simulations

Freedman and Berk (2008) (hereafter, FB) compared weighted and unweighted regression approaches to estimating coefficients in parametric causal models. They demonstrated that propensity score weighting can increase the bias and variance of the estimators relative to unweighted regression, even when the true propensity

score model is known. FB were concerned with how applied researchers were using double robust estimators: to perform structural estimation using parametric models. FB noted that this was far from the original intention of Robins and his collaborators for double robust estimators. Robins et al. were estimating treatment effects (contrasts) and using semiparametric model to perform the estimation. We use the FB simulations to engage exactly that setting: we are estimating treatment effects and we compare the performance of two semiparametric estimators, the TMLE and C-TMLE, relative to alternative estimators. We replicated the original FB simulation study and offer additional simulations based on modifications of their setup. We should note up front that our intention is not to question the take-home point of the FB article: Using double robust estimators to perform structural estimation is fraught with difficulties. We explore different questions, including that of whether nonparametric double robust estimators can be used to estimate treatment effects without the observed instability of the usual estimators relying on the inverse probability of treatment weighting.

We examine the behavior of TMLE, C-TMLE, and A-IPTW, in addition to the WLS and OLS estimators FB consider. The additive treatment effect in the FB simulations is defined nonparametrically as $\Psi(P) = E_P[E_P(Y \mid A = 1, W) - E_P(Y \mid A = 0, W)]$, where n i.i.d. copies of $O = (W, A, Y) \sim P_0$ represents the observed data, with outcome Y, binary treatment assignment A, and covariates W.

FB simulation 1 presents weighted and unweighted linear regression results based on the correct model and two misspecified parametric models, using a data-generating distribution that has conditional treatment assignment probabilities that come close to 0 and 1. We present results from applying each estimator discussed below to FB simulation 1 as well as additional results using modified data-generating distributions that provide additional insight into estimator performance.

21.3.1 Estimators

Given a linear regression model for $E_0(Y \mid A, W)$, the unweighted linear regression (OLS) estimator is obtained with least squares regression, while the weighted linear regression (WLS) estimator is obtained with weighted least squares regression assigning weight $w_i = 1/g_n(A_i \mid W_i)$ to observation $O_i = (W_i, A_i, Y_i)$. The A-IPTW estimator is given by

$$\psi_{A\text{-}IPTW,n} = \frac{1}{n}\sum_{i=1}^{n} \frac{I(A_i = 1) - I(A_i = 0)}{g_n(A_i \mid W_i)}(Y_i - \bar{Q}_n^0(A_i, W_i)) + \bar{Q}_n^0(1, W_i) - \bar{Q}_n^0(0, W_i),$$

where $\bar{Q}_n^0(A, W)$ is a least squares regression estimator of $\bar{Q}_0$, the true conditional mean of Y, given (A, W). The TMLE is a substitution estimator:

$$\psi_{TMLE,n} = \frac{1}{n}\sum_{i=1}^{n}(\bar{Q}_n^*(1, W_i) - \bar{Q}_n^*(0, W_i)),$$

where $\bar{Q}_n^*$ is a targeted estimate of the true regression $\bar{Q}_0(A, W) = E_0(Y \mid A, W)$, obtained by fluctuating the initial estimate, $\bar{Q}_n^0$, in a manner designed to reduce bias in the estimate of the parameter of interest. A logistic fluctuation working model was employed, guaranteeing that the TMLE $\bar{Q}_n^*$ would remain within the bounds $[a, b]$ for the outcome Y, set by design, by the user, or based on the observed outcomes. Since Y is generated from a normal distribution, the bounds were arbitrarily chosen to be the [0.01, 0.99] quantiles of the observed values for Y in each simulated data set, and Y was truncated by these quantiles. The truncated data set can now be viewed as (W, A, Y), with $Y \in [a, b]$ for known bounds $[a, b]$. The TMLE procedure maps $Y \in [a, b]$ into $Y^* = (Y - a)/(b - a) \in [0, 1]$, then regresses Y^* onto A, W to obtain an initial estimator of $E_0(Y^* \mid A, W)$. Since the TMLE involves fluctuating the logit of this initial estimator, the values of the initial estimator were bounded away from 0 and 1 by truncating them from above and below at $\alpha = [0.005, 0.995]$. Recall that a TMLE is defined by a choice of submodel and loss function. Two TMLEs were implemented. In the first TMLE, we used the logistic regression submodel with clever covariate $(2A - 1)/g_n(A \mid W)$ and the quasi-log-likelihood loss function. This TMLE of the additive effect was described in Chap. 7. The second TMLE used the logistic regression submodel with clever covariate $(2A - 1)$, and the weighted quasi-log-likelihood loss function, where the weights are $1/g_n(A_i \mid W_i)$, $i = 1, \ldots, n$. We will denote this latter TMLE with TMLE_w. The latter TMLE is a less aggressive in weighting, and might therefore be more robust under violations of the positivity assumption.

The C-TMLE is also a substitution estimator:

$$\psi_{C-TMLE,n} = \frac{1}{n} \sum_{i=1}^{n} (\bar{Q}_n^*(1, W_i) - \bar{Q}_n^*(0, W_i)).$$

The C-TMLE is described in Chap. 19 and involves building a main term logistic regression estimator of a conditional distribution of A in terms of a set of covariates that are still predictive of Y, after taking into account the initial estimator. Two sets of C-TMLE results were obtained. For the first, labeled C-TMLE in Tables 21.5–21.6, the covariate set W used to create the series of treatment mechanism estimators is restricted to main term covariates. In the second set, labeled C-TMLE(augW), the set of main terms is augmented with four terms corresponding to the propensity score estimate supplied to all other estimators and truncated propensity scores, truncated at level $(p, 1 - p)$, with p set to (0.10, 0.25, 0.50).

21.3.2 Simulations

Two data-generating distributions are defined. For each one, 250 samples of size $n = 1000$ are drawn from the given data-generating distribution. The propensity score g_0 is estimated using the correct probit model for treatment, and the estimator is denoted by g_n. The correct linear regression statistical model includes A, W_1, and

W_2 as main terms. Two increasingly misspecified models are defined, one with A and W_1 as main terms and one with only A as a main term. Estimates of the marginal additive treatment effect are obtained based on an MLE fit of $\bar{Q}_0$ according to these parametric models paired with the MLE g_n of g_0.

Freedman and Berk simulation 1. This simulation replicates FB simulation 1. Both covariates, W_1 and W_2, confound the relationship between treatment and the outcome, so one expects OLS to be biased when the regression model for $\bar{Q}_0$ is misspecified. Incorporating estimated propensity scores should allow the remaining estimators to be unbiased, at the cost of higher variance. Specifically, the data-generating distribution is defined as follows:

$$\begin{aligned}
&Y = 1 + A + W_1 + 2W_2 + U,\ U \sim N(0,1),\\
&P_0(A = 1 \mid W) = \Phi(0.5 + 0.25W_1 + 0.75W_2),\\
&(W_1, W_2) \text{ is bivariate normal, } N(\mu, \Sigma),\\
&\text{with } \mu_1 = 0.5,\ \ \mu_2 = 1,\ \ \Sigma = \begin{bmatrix} 2 & 1 \\ 1 & 1 \end{bmatrix},
\end{aligned}$$

where Φ is the CDF of the standard normal distribution, so that the treatment mechanism conforms to a probit model. These settings lead to finite sample violations of the positivity assumption: conditional treatment probabilities $g_0(1 \mid W) = P_0(A = 1 \mid W)$ range from 0.03 to 0.99995.

Freedman and Berk simulation 2. This simulation was designed to demonstrate that weighting can introduce bias in the estimate of the additive treatment effect, even when the correct propensity score model is known. In this simulation, $P_0(A = 1 \mid W)$ is between 0.0003 and 0.9997. The linear form of the relationships between the covariates and the outcome is unchanged, but the strengths of those relationships are altered to weaken the association between W_1 and W_2, and between W_2 and A, but strengthen the relationships between W_1 and Y and W_2 and Y. As in simulation 2, W_3 is associated with A, but not with the outcome Y. Specifically, the data-generating distribution is defined as follows:

$$\begin{aligned}
&Y = 1 + A + 5W_1 + 10W_2 + U,\ U \sim N(0,1),\\
&P_0(A = 1 \mid W) = \Phi(0.25W_1 + 0.001W_2 + W_3),\\
&(W_1, W_2) \text{ is bivariate normal, } N(\mu, \Sigma), \text{with } \mu_1 = 0.5, \quad \mu_2 = 2, \quad \Sigma = \begin{bmatrix} 0.1 & 1 \\ 1 & 1 \end{bmatrix}.
\end{aligned}$$

21.3.3 Results

OLS, WLS, A-IPTW, TMLE, and C-TMLE were applied to each simulated data set. In all simulations, when the model for the true conditional mean $\bar{Q}_0$ was correctly specified, OLS, the unweighted parametric estimator, had the smallest MSE, but when $\bar{Q}_0$ was misspecified, all other estimators outperformed OLS with respect

to both MSE and bias. Simulation 1 results suggest that TMLE and C-TMLE are more robust than WLS and A-IPTW to practical violations of the positivity assumption. C-TMLE results in simulation 2 demonstrate that performance improves under lack of positivity by using a procedure that estimates only the necessary portion of the treatment mechanism. C-TMLEs' MSE and variance are superior to those of the other estimators that incorporate propensity score estimates. Under extreme misspecification (misspecified model 2) bias is almost entirely removed. Additionally, augmenting the covariate set with truncated propensity scores improves the performance of the C-TMLE. This augmentation confers the greatest benefit when the parametric model for $\bar{Q}_0$ is most severely misspecified. The results for the two simulations are presented in Tables 21.5 and 21.6. It is also of interest to note that for the truncated g_n setting, the TMLE that incorporates g_n in the clever covariate performs better than the TMLE that moves g_n into the weight, while the opposite is true for the unbounded g_n. Since it is theoretically sound to use some bounding, this particular simulation seems to favor the first type of TMLE.

Table 21.5 Freedman and Berk simulation 1, 250 samples of size 1000

	g_n unbounded				g_n bound = (0.025, 0.975)			
	Bias	Var	MSE	RE*	Bias	Var	MSE	RE
Unadj	4.061	0.046	16.538	–	4.061	0.046	16.538	–
Correct model								
OLS	0.010	0.009	0.010	1.000	–	–	–	–
WLS	0.012	0.039	0.039	4.144	0.016	0.024	0.024	2.526
A-IPTW	0.019	0.058	0.059	6.153	0.014	0.017	0.017	1.766
TMLE	0.190	0.475	0.509	53.460	0.019	0.027	0.027	2.834
TMLE_w	0.016	0.047	0.048	4.994	0.015	0.018	0.018	1.909
C-TMLE	0.004	0.014	0.014	1.449	0.013	0.013	0.013	1.410
C-TMLE(augW)	0.011	0.010	0.010	1.092	0.014	0.014	0.014	1.501
Misspecified model 1								
OLS	1.138	0.020	1.314	1.000	–	–	–	–
WLS	0.133	0.115	0.133	0.101	0.295	0.040	0.127	0.096
A-IPTW	0.120	0.344	0.357	0.272	0.433	0.033	0.220	0.167
TMLE	−0.588	0.380	0.724	0.551	−0.001	0.048	0.048	0.037
TMLE_w	0.134	0.177	0.194	0.148	0.359	0.032	0.161	0.123
C-TMLE	0.262	1.516	1.579	1.202	−0.412	0.098	0.267	0.203
C-TMLE(augW)	−0.242	1.068	1.122	0.854	−0.077	0.054	0.060	0.046
Misspecified model 2								
OLS	4.061	0.046	16.538	1.000	–	–	–	–
WLS	0.431	0.660	0.843	0.051	1.070	0.091	1.234	0.075
A-IPTW	0.381	3.039	3.172	0.192	1.507	0.130	2.402	0.145
TMLE	−0.451	1.392	1.590	0.096	−0.132	0.120	0.137	0.008
TMLE_w	0.430	1.226	1.406	0.085	1.260	0.105	1.693	0.102
C-TMLE	1.885	5.358	8.889	0.537	0.456	0.276	0.482	0.029
C-TMLE(augW)	−0.046	0.158	0.160	0.010	0.011	0.063	0.063	0.004

*Relative to OLS estimator using the same model specification

Table 21.6 Freedman and Berk simulation 2, 250 samples of size 1000

	g_n unbounded				g_n bound = (0.025, 0.975)			
	Bias	Var	MSE	RE	Bias	Var	MSE	RE
Unadj	3.022	0.688	9.816	–	3.022	0.688	9.816	–
Correct model								
OLS	0.002	0.004	0.004	1.000	–	–	–	–
WLS	0.002	0.012	0.012	3.175	0.004	0.009	0.009	2.476
A-IPTW	0.004	0.018	0.018	4.694	0.004	0.009	0.009	2.470
TMLE	0.001	0.067	0.067	17.676	0.002	0.011	0.011	3.003
TMLE_w	0.002	0.015	0.015	4.027	0.003	0.009	0.009	2.490
C-TMLE	0.002	0.004	0.004	0.991	0.002	0.004	0.004	0.989
C-TMLE(augW)	0.001	0.004	0.004	1.044	0.001	0.004	0.004	1.059
Misspecified model 1								
OLS	0.024	0.447	0.446	1.000	–	–	–	–
WLS	−0.108	0.500	0.510	1.143	−0.037	0.223	0.224	0.501
A-IPTW	−0.144	0.830	0.847	1.898	−0.037	0.223	0.224	0.502
TMLE	−0.127	1.077	1.089	2.440	−0.053	0.291	0.293	0.656
TMLE_w	−0.134	0.678	0.693	1.553	−0.039	0.227	0.228	0.511
C-TMLE	−0.077	0.050	0.056	0.125	−0.077	0.047	0.053	0.118
C-TMLE(augW)	−0.091	0.042	0.050	0.112	−0.094	0.045	0.054	0.120
Misspecified model 2								
OLS	3.022	0.688	9.816	1.000	–	–	–	–
WLS	−0.077	1.686	1.685	0.172	0.186	0.392	0.425	0.043
A-IPTW	−0.167	3.727	3.740	0.381	0.232	0.406	0.459	0.047
TMLE	−0.940	1.357	2.235	0.228	−0.294	0.555	0.639	0.065
TMLE_w	−0.120	2.181	2.187	0.223	0.180	0.400	0.430	0.044
C-TMLE	0.002	0.073	0.073	0.007	−0.005	0.021	0.021	0.002
C-TMLE(augW)	−0.049	0.073	0.075	0.008	−0.033	0.045	0.046	0.005

The C-TMLE outperforms all of the other estimators except for when the correct OLS model is used. And the performance of the C-TMLE is never poor in any simulation. The usual instability problems of weighted estimators has been minimized. The combination of properties in the C-TMLE proved especially robust in these Monte Carlos: it is a double robust (asymptotically efficient) substitution estimator that respects global constraints, makes use of a logistic fluctuation to respect the bounds even in finite samples, and performs internal collaborative estimation for g_0.

Domain knowledge can be incorporated into both stages of the TMLE and C-TMLE procedures. One example is the use of the augmented covariate set when the true treatment assignment mechanism is known. The strength of this approach is most clearly illustrated in simulation 2 with $\bar{Q}_0$ modeled with misspecified model 2, where the right thing to do is adjust for all covariates, yet that causes strong positivity violations. In this case, the inclusion of truncated propensity scores in W offered a more refined choice beyond simply including or excluding an entire covariate. These additional terms can be helpful in situations where including a particular co-

variate causes an positivity violation, but in fact, experimentation is lacking in only some portion of the covariate values.

21.4 Discussion

Researchers spend too little time on design and too much time on analysis in an attempt to overcome design defects. Sometimes – and in some fields, such as the social sciences, often – the correct answer is that the data at hand cannot answer the research question. Often new data must be gathered with a better design, ideally a design in which the researcher exploits natural or intentional variation to mitigate confounding instead of having to make a selection on observables assumption.

An essential goal of a scientific study is objectivity. Relying on an estimation strategy where one adjusts the model specification or estimator after one has observed estimated treatment effects cannot be considered objective. However, this objectivity is fully addressed by any estimator of the target parameter that is a priori specified, or at most influenced by ancillary statistics, so that the pursuit of objectivity itself should not limit the choice of estimators. The utilization of an a priori specified machine (and, specifically, super) learning algorithm to perform the modeling helps to mitigate the data-snooping concerns: the estimation procedure is fully specified before the analyst observes any final outcome data or estimated treatment effects. Having resolved the concern for objectivity, the remaining concern then centers on the instability of most double robust estimators when the data are sparse. C-TMLE is more stable than the other estimators considered here, and the TMLE and C-TMLE with logistic fluctuations perform well in these simulations.

The TMLE and C-TMLE and their accompanying technology, such as the super learner, are powerful and promising tools that overcome some of the common objections to double robust estimators. Demonstrating that the TMLE and C-TMLE perform well in general when the positivity assumption is violated is difficult because sparsity is a finite sample concern, and the efficiency and double robustness of TMLE and C-TMLE are asymptotic statistical properties, but the fact that these estimators are also *substitution* estimators (i.e., obtained by plugging an estimator of the data-generating distribution into the statistical model) explains the observed robustness. In particular, a substitution estimator puts bounds on the influence of one observation fully implied by the statistical model and the target parameter as a mapping on that statistical model. We hope that by showing that these estimators perform well in simulations created by *other* researchers for the purposes of showing the weaknesses of double robust estimators, we provide probative evidence in support of TMLE and C-TMLE. Indeed, we also extended the original simulations to make the estimation problems more challenging. Of course, much can happen in finite samples, and we look forward to exploring how these estimators perform in other settings. Of particular interest are applications of this technology to applied problems.