University of California, Berkeley
U.C. Berkeley Division of Biostatistics Working Paper Series

Year 2007 *Paper* 215

Covariate Adjustment in Randomized Trials with Binary Outcomes: Targeted Maximum Likelihood Estimation

Kelly L. Moore* Mark J. van der Laan†

*Division of Biostatistics, School of Public Health, University of California, Berkeley, klmoore@stat.berkeley.edu

†Division of Biostatistics, School of Public Health, University of California, Berkeley, laan@berkeley.edu

http://biostats.bepress.com/ucbbiostat/paper215

Covariate Adjustment in Randomized Trials with Binary Outcomes: Targeted Maximum Likelihood Estimation

Kelly L. Moore and Mark J. van der Laan

Abstract

Covariate adjustment using linear models for continuous outcomes in randomized trials has been shown to increase efficiency and power over the unadjusted method in estimating the marginal effect of treatment. However, for binary outcomes, investigators generally rely on the unadjusted estimate as the literature indicates that covariate-adjusted estimates based on logistic regression models are less efficient. The crucial step that has been missing when adjusting for covariates is that one must integrate/average the adjusted estimate over those covariates in order to obtain the marginal effect. We apply the method of targeted maximum likelihood estimation (MLE), as presented in van der Laan and Rubin (2006), to obtain estimators for the marginal effect using covariate adjustment for binary outcomes. We show that the covariate adjustment in randomized trials using logistic regression models can be mapped, by averaging over the covariate(s), to obtain a fully robust and efficient estimator of the marginal effect, which equals the targeted maximum likelihood estimator (MLE). We present simulation studies that show the targeted MLE increases efficiency and power over the unadjusted method, particularly for smaller sample sizes, even when the regression model is mis-specified.

1 Introduction

Suppose we observe n independent and identically distributed observations of the random vector $O = (W, A, Y) \sim p_0$, where W is a vector of baseline covariates, A is the treatment of interest and $Y = \{0, 1\}$ is the binary outcome of interest, and p_0 denotes the density of O. Causal effects are based on a hypothetical full data structure $X = ((Y_a : a \in \mathcal{A}), W)$ containing the entire collection of counterfactual or potential outcomes Y_a for a ranging over the set of all possible treatments $\mathcal{A}$. The observed data structure O only contains a single counterfactual outcome $Y = Y(A)$ corresponding to the treatment that the subject received. The observed data $O = (W, A, Y \equiv Y(A))$ is thus a missing data structure on X with missingness variable A. We denote the conditional probability distribution of treatment A by $g_0(a|X) \equiv P(A = a|X)$. The randomization assumption or coarsening at random assumption states that A is conditionally independent of the full data X given W, $g_0(A|X) = g_0(A|W)$. In a randomized trial in which treatment is assigned completely at random, we have $g_0(A|X) = g_0(A)$. For the sake of presentation, we assume the treatment A is binary and that A is completely randomized as in a typical randomized trial, but our methods are presented so that it is clear how our estimators generalize to observational studies or randomized trials in which $g_0(A|W)$ is known. In the binary A case, $g_0(1) = p(A = 1) = \delta_0$ and $g_0(0) = p(A = 0) = 1 - \delta_0$ and n_1 the number of subjects in treatment group 1 and n_0 the number of subjects in treatment group 0, and $n = n_1 + n_0$. The quantity of interest is causal effect of treatment A on Y, which, for example, can be defined as the risk difference $\psi = E(Y_1) - E(Y_0)$, where Y_1 and Y_0 are the counterfactual outcomes under treatments 1 and 0 respectively. This quantity is typically estimated in randomized trials with the unadjusted estimate

$$\hat{\psi}_1 = \hat{\mu}_1 - \hat{\mu}_0$$

where $\hat{\mu}_1 = \frac{1}{n}\sum_{i=1}^{n} I(A_i = 1)Y_i$ and $\hat{\mu}_0 = \frac{1}{n}\sum_{i=1}^{n} I(A_i = 0)Y_i$. An adjusted effect is also sometimes obtained,

$$\hat{\psi}_W = \hat{P}(Y = 1|A = 1, W) - \hat{P}(Y = 1|A = 0, W).$$

Adjusting for baseline covariates and the issues involved has been discussed in Pocock et al. (2002). Although it has been recognized, at least for linear models, i.e. continuous outcomes, that adjusting for covariates increases the precision of the estimate of the marginal causal effect of treatment, investigators are still resistant to adjusting in logistic models and often rely on the unadjusted estimate. This generally appears to be due to confusion as to how to select the covariates and and how to adjust for them (Pocock et al. 2002). In addition, there is a concern that if data-adaptive procedures are used to select the model for $P(Y = 1|A, W)$ that investigators will be tempted to select the model that provides the most favorable results. However, we recommend that as long as the procedure is determined a priori then we can avoid this latter issue. Thus, a black box type data-adaptive procedure, e.g. forward selection, can still be

applied as long as the algorithm and candidate covariates are specified a priori. Adjusting for covariates with main terms in linear models, referred to as analysis of covariance (ANCOVA) in randomized trial literature, for the purpose of estimation of the marginal causal effect has been limited to no interaction terms with treatment. When there is such an interaction term, it is often not clear in the literature on analysis of randomized trial data how one uses this conditional model to obtain a marginal effect. However, even in the absence of the interaction term, the increase in precision has not been observed for non-linear models such as the logistic model. In fact, it has actually been reported that the estimates are not in fact made more precise for logistic models. The crucial step that has been missing when the parameter of interest is the *marginal* causal effect of A on Y, is that when adjusting for covariates W, one must integrate/average the adjusted estimate over those W in order to obtain a marginal effect estimate that is comparable to the unadjusted effect estimate $\hat{\psi}_1$. This method of averaging over W has been referred to as the G-computation formula and is often applied in observational studies when the treatment or exposure has not been assigned randomly (Robins (1986) and Robins (1987)). We show that with this additional step of averaging over W, even when the outcome is binary, and even if the regression model is misspecified, we obtain a more efficient estimate in the randomized trial setting. Such an approach allows for interactions between A and W in the model for $P(Y = 1|A, W)$ while still obtaining a marginal effect. We note that the conditional effect may be the parameter of interest in some studies, for example the effect of a drug conditional on age, and thus the investigator does not want to average over age. In this paper we focus only on the marginal effect and using the covariates W to obtain the most efficient (precise) estimate of this marginal causal effect in a nonparametric model. We apply the method of targeted maximum likelihood estimation (MLE), as presented in van der Laan and Rubin (2006), to obtain estimators for the marginal effect using covariate adjustment for binary outcomes. This general targeted MLE methodology applies to any estimation problem. In this article we apply it to the risk difference, relative risk and odds ratio, in the context of a randomized trial. Targeted MLE was purposefully named in that maximum likelihood estimators aim for trade-off between bias and variance for the whole density, while the targeted MLE carries out a bias reduction specifically *taylored* for the parameter of interest. Substitution estimators based on standard MLE are often biased with respect to the parameter of interest and do not always converge at a parametric rate. On the other hand, the targeted MLE maps a density estimator (e.g., MLE) into a targeted maximum likelihood estimator (at parameter of interest) so that the corresponding substitution estimator is double robust and locally efficient. That is, this estimator in the randomized trial setting is always consistent and asymptotically linear even when the initial regression estimator for $P(Y|A, W)$ is mis-specified, and is even nonparametrically efficient if the initial estimator is consistent. The general algorithm provided in van der Laan and Rubin (2006) is to start with initial density estimator, then create a parametric model with parameter ϵ through this given initial density estimator whose scores at $\epsilon = 0$ include the components of the efficient influence

curve of the parameter of interest at the given density estimator. It estimates ϵ with MLE of this parametric model and finally updates the new density estimator as the corresponding fluctuation of the given initial density estimator. The algorithm can be iterated until convergence. However in many examples convergence is achieved in a single step as is the case for the examples in this paper. We apply this approach to the estimation of marginal treatment effects including the risk difference, relative risk and odds ratio. The targeted maximum likelihood estimator is a very practically attractive procedure since it can be achieved by simply adding a covariate to an initial estimate of the regression $P(Y = 1|A, W)$. The corresponding coefficient ϵ for this new covariate can be estimated with standard software and thus has a straightforward implementation. We show that for the logistic regression model for $P(Y = 1|A, W)$, that this covariate is none other than a linear combination of the treatment variable A so that it follows that the targeted MLE coincides with the standard G-computation ML estimator. This is not always true as we show that these 2 estimators differ when the treatment mechanism is estimated from the data, which results in an additional efficiency gain. In van der Laan, Rubin (2006) we appeal to estimating function methodology (van der Laan, Robins (2002)) and observe that since the targeted MLE solves the efficient influence curve estimating equation it is double robust and (locally) efficient. That is, the targeted MLE is always consistent and asymptotically linear (thus the standardized estimator is asymptotically normally distributed with specified variance), even if the initial estimate for $P(Y = 1|A, W)$ is misspecified. In the case that the initial estimate for $P(Y = 1|A, W)$ is asymptotically consistent the targeted MLE is asymptotically efficient for the nonparametric model. In section 2 we provide a brief overview of methods for covariate adjustment that have been proposed in literature. In section 3 we present the targeted maximum likelihood estimators for three marginal variable importance parameters: the risk difference, relative risk and odds ratio. We show that for each of these three parameters, using a logistic regression model, the targeted MLE is achieved in a single step. We also provide an alternative to the logistic regression model for the relative risk parameter that is the relative risk regression model and provide the corresponding targeted MLE estimator. We also address missing data on the outcome of covariates, and estimation of the treatment mechanism. Section 4 provides testing and inference for the targeted MLE. In section 5 we present simulation studies that demonstrate the performance of the targeted MLE. Finally we conclude with a discussion in section 6.

2 Current Methods for Obtaining Covariate-Adjusted Estimates

Suppose we observe $O = (W, A, Y)$ as above except the outcome Y is now continuous. Let the parameter of interest be the marginal effect of A on Y, $\psi = E(Y_1) - E(Y_0)$. For a continuous outcome Y, $Q(A, W) = E(Y|A = 1, W)$

is typically obtained using a linear regression model such as,

$$\hat{Q}(A,W) = \hat{\beta}_0 + \hat{\beta}_1 A + \hat{\beta}_2 W.$$

In this setting, $\hat{\beta}_1$ coincides with and has been shown to be at least as precise as the unadjusted estimate $\hat{\psi}_1$. In particular, the increase in precision occurs when the correlation between the covariate(s) and outcome is strong (Assmann et al. 2002). However, when $Q(A,W)$ is estimated as

$$\hat{Q}(A,W) = \hat{\beta}_0 + \hat{\beta}_1 A + \hat{\beta}_2 W + \hat{\beta}_3 AW,$$

then $\hat{\beta}_1$ no longer coincides with $\hat{\psi}_1$. In this case, to obtain the *marginal* effect, one must integrate out or average over the covariate(s) W. Robins (1986) and Robins (1987) introduced the G-computation estimator that does indeed average over W and thus give a marginal effect,

$$\hat{\psi}_{Gcomp} = \frac{1}{n}\sum_{i=1}^{n} \hat{Q}(1,W_i) - \hat{Q}(0,W_i).$$

When $\hat{Q}(A,W)$ is estimated with a linear model, and it does not contain any interaction terms, then $\hat{\psi}_{Gcomp} = \hat{\beta}_1$. The G-computation estimator is not limited to a linear model for $Q(A,W)$ when estimating the treatment effect, for example, when the outcome is binary, one could use a logistic regression model to estimate $Q(A,W)$ and use the G-computation formula to obtain the estimated risk difference. However, even in the absence of interaction terms, $\hat{\psi}_{Gcomp}$ is not necessarily equivalent to the estimate obtained from the logistic regression model. Based on estimating function methodology, the Double Robust (DR) estimator has been provided in Robins (2000), Robins and Rotnitzky (2001) and Neugebauer and van der Laan (2002), van der Laan and Robins (2002). Consistency of the DR estimator relies on consistent estimation of the treatment mechanism *or* the model for $Q(A,W)$. When the treatment is randomized, as in a randomized trial, the treatment mechanism is always known and thus the DR estimator is always consistent, i.e. even when $Q(A,W)$ is mis-specified. Scharfstein et al. (1999, p. $1140-1141$) showed that to obtain a DR estimate, one can update $Q(A,W)$ by adding the 2-dimensional covariate $(\frac{I(A=1)}{g(1|W)}, \frac{I(A=0)}{g(0|W)})$. Note that in the randomized trial setting, $g(1|W) = \delta$ and $g(0|W) = 1-\delta$. In section 3.4, under the framework of targeted MLE, we also propose adding these two covariates, the first for $P(Y_1 = 1)$ and one for $P(Y_0 = 1)$ so that any function of these two parameters is estimated in a targeted manner. The resulting estimator that targets the 2-dimensional parameter equals the proposed estimator of Scharfstein et al. (1999, p. $1140 - 1141$). Bang and Robins (2005) indicate that when the initial model for $Q(A,W)$ is correct, then one can obtain a more efficient DR estimate by adding the 1-dimensional covariate $\frac{I(A=1)}{g(1|W)} + \frac{I(A=0)}{g(0|W)}$. We derive this 1-dimensional covariate under the framework of targeted MLE in section 3.1, targeting the parameter of interest the risk difference. Note that this covariate differs when the parameter of interest is the relative risk or odds

ratio as shown in sections 3.2 and 3.3. In section 3.1.1 we provide the relation between the DR, targeted MLE and G-computation estimator and the circumstances in which they coincide. Tsiatis et al. (2006) applies this DR estimator for the marginal effect where the authors recommend estimating two regression models separately: $Q_1(1, W) = E(Y|A = 1, W)$ is obtained using only the subpopulation of individuals for whom $A = 1$ and $Q_2(0, W) = E(Y|A = 0, W)$ is obtained using only the subpopulation of individuals for whom $A = 0$. This was proposed so that two different analysts could independently select these models to prevent the analysts from selecting the model providing the most favorable results. Another possibility is to select one model $Q(A, W) = E(Y|A, W)$ using the whole sample pooled together. When the procedure for selecting $Q(A, W)$ is specified a priori this additional step of estimating $Q_1(1, W)$ and $Q_2(0, W)$ is not necessary. The method provided by Tsiatis et al. (2006) is limited to when the parameter of interest of the marginal effect $E(Y_0) - E(Y_1)$. However, when the outcome is binary, investigators are often also interested in not only the risk difference $E(Y_0) - E(Y_1) = P(Y_1 = 1) - P(Y_0 = 1)$, but the relative risk and odds ratios. Covariate adjustment in logistic regression models for binary outcomes has been studied in literature. However it does not appear that any method for covariate adjustment has been proposed to obtain *marginal* estimates for such parameters. Thus, current applications of logistic regression models provide conditional effects. These conditional models have been shown to *reduce* precision in the estimated effect. Robinson and Jewell (1991) observed that adjusting for covariates in logistic regression models leads to an increase in power due to the fact that estimates of the treatment effect in the conditional logistic models are further away from the null even though standard errors were larger for the adjusted effects. Hernández et al. (2004) also demonstrated this fact using using simulation studies and observed that the increase in power was related to the correlation between the covariate and the outcome. The simulations included only a single covariate and no interactions between the covariate and treatment. Assmann et al. (2000) also indicated similar results in logistic regression models in that odds ratios were generally further away from the null but the standard errors were larger than the unadjusted estimates. It appears that in general, when adjusting for covariates in a logistic regression model, the standard error provided by the software, i.e. standard maximum likelihood procedures, is the standard error used by the investigator although it is often not explicitly stated (van der Horst et al. (1997), Randolph et al. (2002), Belda et al. (2005) , Frasure-Smith et al. (1997)). When adjusting for covariates in randomized trials using logistic regression, often the investigator is interested in a conditional effect identified by continuous covariates in which case this may be an appropriate approach. We focus on the targeted MLE method for covariate adjustment that provides inference for the marginal (unconditional) effect. However, note that this method can be applied to different subgroups defined by categorical or discrete valued covariates by simple stratification.

3 Targeted Maximum Likelihood Estimation of Marginal Variable Importance: Risk Difference, Relative Risk and Odds Ratio

In this section we present the targeted MLE method for adjusting for covariates when the outcome is binary with the following 3 parameters: risk difference, relative risk and odds ratio.

3.1 Risk Difference

We now provide the targeted MLE for the risk difference $P(Y_1 = 1) - P(Y_0 = 1)$. Let $O = (W, A, Y) \sim p_0$ and $\mathcal{M}$ be the class of all densities of O with respect to an appropriate dominating measure: so $\mathcal{M}$ is nonparametric up to possible smoothness conditions. Consider this non-parametric model for p_0 and let

$$P_0 \to \Psi(p_0) = E_{p_0}(P(Y|A = 1, W) - P(Y|A = 0, W))$$

be the parameter of interest. This parameter is pathwise differentiable at p_0 with efficient influence curve,

$$\begin{aligned} D(p_0) &= \frac{I(A=1)}{\delta_0}(Y - Q_0(1, W)) - \frac{I(A=0)}{(1-\delta_0)}(Y - Q_0(0, W)) + \\ &\quad + Q_0(1, W) - Q_0(0, W) - \Psi(p_0) \end{aligned}$$

where $Q_0(A, W) = P(Y = 1|A, W)$ and $\delta_0 = P(A = 1)$ (see e.g., van der Laan, Robins, 2002). Since the model is non-parametric, this is also the only influence curve. Following the strategy of van der Laan and Rubin (2006), the efficient influence curve $D(p_0)$ can be decomposed as,

$$\begin{aligned} D(p_0) &= D(p_0) - E(D(p_0)|A, W) + E(D(p_0)|A, W) - E(D(p_0)|W) + \\ &\quad + E(D(p_0)|W) - E(D(p_0)) \end{aligned}$$

Let, $D_1(p_0) = D(p_0) - E(D(p_0)|A, W)$, $D_2(p_0) = E(D(p_0)|A, W) - E(D(p_0)|W)$ and $D_3(p_0) = E(D(p_0)|A, W) - E(D(p_0))$. Then, $D_1(p_0)$ is a score for $p(Y|A, W)$, $D_2(p_0)$ is a score for $g_0(A|W)$ and $D_3(p_0)$ is a score for the marginal probability distribution $p(W)$ of W. Note that in this randomized trial setting, $g_0(A|W) = g_0(A) = \delta_0^A(1 - \delta_0)^{(1-A)}$.

Consider an initial density estimator $\hat{p}^0$ of the density p_0 of O identified by a regression fit $\hat{Q}^0(A, W)$, marginal distribution of A identified by $\hat{\delta} = \frac{1}{n}\sum_{i=1}^n A_i$, the marginal distribution of W being the empirical probability distribution of $W_1, ..., W_n$, and A being independent of W. Since Y is binary, we have the following density,

$$\hat{p}^0(Y|A, W) = (\hat{Q}^0(A, W))^Y (1 - \hat{Q}^0(A, W))^{1-Y}$$

where,

$$\hat{Q}^0(A,W) = \frac{1}{1+\exp -\hat{m}^0(A,W)}$$

for some function $\hat{m}^0$. Now, consider the parametric submodel through $\hat{p}^0$ indexed by parameter ϵ,

$$\hat{p}^0(\epsilon)(Y|A,W) = (\hat{Q}^0(\epsilon)(A,W))^Y(1-\hat{Q}^0(\epsilon)(A,W))^{1-Y}$$

where $\hat{Q}^0(\epsilon)(A,W)$ is given by the logistic regression model,

$$\hat{Q}^0(\epsilon)(A,W) = \frac{1}{1+\exp -(\hat{m}^0(A,W)+\epsilon h(A,W))}$$

with an extra covariate $h(A,W)$, which needs to be chosen so that the score of ϵ at $\epsilon = 0$ includes the efficient influence curve component $D_1(p^0)$ (see van der Laan, Rubin, 2006). The required choice h will be specified below. We estimate ϵ with the maximum likelihood estimator $\hat{\epsilon} = \arg\max_\epsilon \sum_{i=1}^n \log \hat{Q}^0(\epsilon)(A_i,W_i)$. The score for this logistic regression model at $\epsilon = 0$ is given by,

$$\frac{d}{d\epsilon_1}\log p^0(\epsilon)(A,W)\Big|_{\epsilon=0} = h(A,W)(Y-\hat{Q}^0(A,W))$$

We now set the score equal to the part of the efficient IC for $p(Y|A,W)$, that is D_1, at $\hat{p}^0$ to obtain,

$$h(A,W)(Y-\hat{Q}^0(A,W)) = (Y-\hat{Q}^0(A,W))\left(\frac{I(A=1)}{\hat{\delta}} - \frac{I(A=0)}{(1-\hat{\delta})}\right).$$

This equality in $h(A,W)$ is solved by

$$h(A,W) = \frac{I(A=1)}{\hat{\delta}} - \frac{I(A=0)}{(1-\hat{\delta})}.$$

Thus, the covariate that is added to the logistic regression model $\hat{Q}^0(A,W)$ is none other than a linear combination of A and an intercept only. Thus, if $\hat{m}^0(A,W)$ includes the main term A and the intercept, then $\hat{\epsilon} = 0$, and the targeted MLE for $Q_0(A,W)$ is given by $\hat{Q}^0(A,W)$ itself. In other words, the targeted MLE for ψ_0 is given by the standard G-computation estimator

$$\hat{\psi}_{RD-tMLE} = \frac{1}{n}\sum_{i=1}^n \hat{Q}^0(1,W_i) - \hat{Q}^0(0,W_i).$$

3.1.1 Relation between Targeted MLE, DR and G-computation Estimators

The efficient influence curve $D(p_0)$ can be represented as an estimating function in ψ indexed by Q and g, $D(p_0) = D(Q_0, g_0, \Psi(p_0))$. In this randomized trial

setting, $g_0 = \delta_0^A(1-\delta)^{1-A}$. The DR estimate is the solution to the corresponding estimating equation in ψ, $\frac{1}{n}\sum_{i=1}^n D(\hat{Q}^0(A_i, W_i), \hat{\delta}, \psi) = 0$ and is given by,

$$\begin{aligned}\hat{\psi}_{DR} &= \frac{1}{n}\sum_{i=1}^n \frac{I(A_i=1)}{\hat{\delta}}(Y_i - \hat{Q}^0(1, W_i)) - \frac{1}{n}\sum_{i=1}^n \frac{I(A_i=0)}{1-\hat{\delta}}(Y_i - \hat{Q}^0(0, W_i)) + \\ &+ \frac{1}{n}\sum_{i=1}^n \hat{Q}^0(1, W_i) - \frac{1}{n}\sum_{i=1}^n \hat{Q}^0(0, W_i),\end{aligned}$$

where $\hat{\delta} = \frac{1}{n}\sum_{i=1}^n A_i$. In the logistic regression fit, $\log(\frac{\hat{Q}(A,W)}{1-\hat{Q}(A,W)}) = \hat{\alpha}X$, where $X = (1, A, W)$, the MLE $\hat{\alpha}$ solves the score equations given by,

$$0 = \sum_{i=1}^n X_{ij}(Y_i - \hat{Q}(A_i, W_i)),$$

for $j = 1, ..., p$. The linear span of scores includes the covariate,

$$x_j = \frac{I(A=1)}{\hat{\delta}} - \frac{I(A=0)}{1-\hat{\delta}},$$

when A and an intercept are included in X. Thus, it follows that

$$0 = \frac{1}{n}\sum_{i=1}^n \frac{I(A_i=1)}{\hat{\delta}}(Y_i - \hat{Q}^0(1, W_i)) - \frac{1}{n}\sum_{i=1}^n \frac{I(A_i=0)}{1-\hat{\delta}}(Y_i - \hat{Q}^0(0, W_i)).$$

Hence,

$$\hat{\psi}_{DR} = \frac{1}{n}\sum_{i=1}^n \hat{Q}(1, W_i) - \frac{1}{n}\sum_{i=1}^n \hat{Q}(0, W_i) = \hat{\psi}_{Gcomp} = \hat{\psi}_{RD-tMLE}$$

Thus in this quite general scenario, we have that the double robust estimator, the G-computation estimator, and the targeted MLE, all reduce to the same estimator.

3.2 Relative Risk

We now consider the parameter

$$P_0 \to \Psi(p_0) = \frac{E_{p_0}(P(Y|A=1, W))}{E_{p_0}(P(Y|A=0, W)))} = \frac{\mu_1}{\mu_0}$$

Note that under the assumptions listed above for the risk difference, this parameter can be interpreted as the causal relative risk, $\psi_0 = \frac{E(Y_1)}{E(Y_0)}$.

We can derive the efficient influence curve of this parameter using the delta method since we know the efficient influence curve for μ_1 and μ_0. Let $a = \mu_0$

and $b = \mu_1$, so $\psi_0 = \frac{b}{a}$. Then, $\frac{d}{db}\left(\frac{b}{a}\right) = \frac{1}{a}$ and $\frac{d}{da}\left(\frac{b}{a}\right) = -\left(\frac{b}{a^2}\right)$. Thus, the efficient influence curve is given by,

$$\begin{aligned} D(p_0) &= \frac{1}{\mu_0}\left(\frac{I(A=1)}{\delta_0}(Y - Q_0(1,W)) + Q_0(1,W) - \mu_1\right) - \\ &\quad -\frac{\mu_1}{\mu_0^2}\left(\frac{I(A=0)}{(1-\delta_0)}(Y - Q_0(0,W)) + Q_0(0,W) - \mu_0\right) \\ &= \frac{1}{\mu_0}\left(\frac{I(A=1)}{\delta_0}(Y - Q_0(1,W)) + Q_0(1,W)\right) - \\ &\quad \frac{\mu_1}{\mu_0^2}\left(\frac{I(A=0)}{(1-\delta_0)}(Y - Q_0(0,W)) + Q_0(0,W)\right) \end{aligned}$$

We consider two models for the targeted MLE of the relative risk: logistic regression model and the relative risk regression model. In order to find the covariate $h(A,W)$ that is added to the regression model, we note the following equality given in van der Laan and Robins 2002,

$$V(Y,A,W) = (V(1,A,W) - V(0,A,W))(Y - Q(A,W)), \tag{1}$$

if V is a function with conditional mean 0 given A and W. We apply this equality to $D(p_0) = V(Y,A,W)$ to obtain $h(A,W)$.

3.2.1 Submodel 1: Logistic Regression Model

Let $\hat{p}^0(\epsilon_1)$ be the logistic regression fit with an extra covariate extension $\epsilon_1 h(A,W)$. Based on (1) we can immediately observe that the covariate $h(A,W)$ added to the logistic regression is $V(1,A,W) - V(0,A,W)$ since,

$$\begin{aligned} \frac{d}{d\epsilon}\log \hat{p}^0(\epsilon)(A,W)\Big|_{\epsilon=0} &= h(A,W)(Y - \hat{Q}^0(A,W)) \\ &= (V(1,A,W) - V(0,A,W))(Y - \hat{Q}^0(A,W)) \end{aligned}$$

Thus, evaluating $D(\hat{p}_0)$ at $Y = 1$ and $Y = 0$ gives,

$$h(A,W) = \frac{1}{\mu_0}\frac{I(A=1)}{\hat{\delta}} - \frac{\mu_1}{\mu_0^2}\frac{I(A=0)}{(1-\hat{\delta})}.$$

Again, as in the risk difference, the covariate that is added to $\hat{Q}^0(A,W)$ is a function of A only and thus $\hat{\epsilon} = 0$ and the targeted MLE for $Q_0(A,W)$ is given by $\hat{Q}^0(A,W)$. The targeted MLE for the relative risk is given by,

$$\hat{\psi}_{RR-tMLE} = \frac{\frac{1}{n}\sum_{i=1}^n \hat{Q}^0(1,W_i)}{\frac{1}{n}\sum_{i=1}^n \hat{Q}^0(0,W_i)}.$$

3.2.2 Submodel 2: Relative Risk Regression

As an alternative to using a logistic fit $Q^0(A,W)$ for $Q(A,W)$, we can instead use a relative risk regression fit,

$$\log(\hat{Q}(A,W)) = \hat{m}(A,W),$$

and find the corresponding targeted MLE. Consider now the parametric submodel $\hat{p}^0$ indexed by parameter ϵ,

$$\hat{p}^0(\epsilon)(Y|A,W) = (\hat{Q}^0(\epsilon)(A,W))^Y(1-\hat{Q}^0(\epsilon)(A,W))^{1-Y}$$

where $\hat{Q}^0(\epsilon)(A,W)$ is given by the relative risk regression model,

$$\log(\hat{Q}^0)(\epsilon)(A,W) = \hat{m}^0(A,W) + \epsilon h(A,W).$$

The score for this model evaluated at $\epsilon = 0$ is given by,

$$\frac{d}{d\epsilon}\log\hat{p}^0(\epsilon)(A,W)\Big|_{\epsilon=0} = \frac{h(A,W)}{1-\hat{Q}^0(A,W)}(Y-\hat{Q}^0(A,W)),$$

and it follows that the covariate added to logistic regression model to obtain the targeted MLE is given by,

$$h(A,W) = \left(\frac{1}{\mu_0}\frac{I(A=1)}{\hat{\delta}} - \frac{\mu_1}{\mu_0^2}\frac{I(A=0)}{(1-\hat{\delta})}\right)(1-\hat{Q}^0(A,W)).$$

Now $\hat{\epsilon} = \arg\max_\epsilon \sum_{i=1}^n \log\hat{Q}^0(\epsilon)(A_i,W_i)$ can be estimated in practice by fitting a relative risk regression in $\hat{m}^0(A,W)$ and $h(A,W)$, fixing the coefficient in front of $\hat{m}^0(A,W)$ to 1 and the intercept to 0. The resulting coefficient for $h(A,W)$ is $\hat{\epsilon}$. In this case, the covariate is no longer simply a function of A and thus $\hat{\epsilon}$ does not necessarily equal 0 and the targeted MLE is no longer achieved in one step but rather iteratively. Now $\hat{Q}^k(A,W)$ is updated as,

$$\log(\hat{Q}^{k+1}(A,W)) = \hat{m}^k(A,W) + \hat{\epsilon}h^k(A,W),$$

setting $k = k+1$ and one iterates this updating step.

3.3 Odds Ratio

We now consider the parameter

$$P_0 \to \Psi(p_0) = \frac{E_{p_0}(P(Y|A=1,W))/(1-E_{p_0}(P(Y|A=1,W)))}{E_{p_0}(P(Y|A=0,W))/(1-E_{p_0}(P(Y|A=0,W)))} = \frac{\mu_1/(1-\mu_1)}{\mu_0/(1-\mu_0)}$$

Note that under the assumptions listed above for the risk difference, this parameter can be interpreted as the causal odds ratio, $\frac{E(Y_1)/(1-E(Y_1))}{E(Y_0)/(1-E(Y_0))}$. Again, applying the delta method we can obtain the efficient influence curve for this parameter.

Let $a = \mu_0$ and $b = \mu_1$, so $\psi = \frac{b/(1-b)}{a/(1-a)}$. Then, $\frac{d}{db}\left(\frac{b/(1-b)}{a/(1-a)}\right) = \frac{(1-a)}{a(1-b)^2}$ and $\frac{d}{da}\left(\frac{b/(1-b)}{a/(1-a)}\right) = -\left(\frac{b}{a^2(1-b)}\right)$. Thus, the efficient influence curve is given by,

$$\begin{aligned} D(p_0) &= \frac{1-\mu_0}{\mu_0(1-\mu_1)^2}\left(\frac{I(A=1)}{\delta_0}(Y - Q_0(1,W)) + Q_0(1,W) - \mu_1\right) - \\ &\quad - \frac{\mu_1}{(\mu_0)^2(1-\mu_1)}\left(\frac{I(A=0)}{(1-\delta_0)}(Y - Q_0(0,W)) + Q_0(0,W) - \mu_0\right) \end{aligned}$$

Applying equality (1) to $D(\hat{p}^0)$, we obtain,

$$h(A,W) = \frac{(1-\mu_0)}{\mu_0(1-\mu_1)^2}\frac{I(A=1)}{\hat{\delta}} - \frac{\mu_1}{\mu_0^2(1-\mu_1)}\frac{I(A=0)}{(1-\hat{\delta})}$$

Again, the covariate that is added to the logistic regression model $\hat{Q}^0(A,W)$ is none other than a function of A only and thus $\hat{\epsilon} = 0$ and the targeted MLE for $Q_0(A,W)$ is given by $\hat{Q}^0(A,W)$. Thus, the targeted MLE for ψ is given by,

$$\hat{\psi}_{OR-tMLE} = \frac{\left(\frac{1}{n}\sum_{i=1}^n \hat{Q}^0(1,W_i)\right) / \left(1 - \frac{1}{n}\sum_{i=1}^n \hat{Q}^0(1,W_i)\right)}{\left(\frac{1}{n}\sum_{i=1}^n \hat{Q}^0(0,W_i)\right) / \left(1 - \frac{1}{n}\sum_{i=1}^n \hat{Q}^0(0,W_i)\right)}.$$

3.4 Targeted MLE for the two treatment specific means, and thereby for all parameters.

Consider the odds ratio, as an example. An alternative for targeting the odds ratio is to simultaneously target both μ_1 and μ_0 and simply evaluate the odds ratio from the targeted MLEs of μ_1 and μ_0. This is a straightforward approach where 2 covariate extensions are added to the logistic fit $\hat{Q}^0$,

$$h_1(A,W) = \epsilon_1 \frac{I(A=1)}{\hat{\delta}},$$

and,

$$h_2(A,W) = \epsilon_2 \frac{I(A=0)}{(1-\hat{\delta})}.$$

Again, if the initial logistic regression fit already includes an intercept and main term A, then $\hat{\epsilon} = 0$ so that this targeted MLE $\hat{Q} = \hat{Q}^0(\hat{\epsilon}) = \hat{Q}^0$ is not updated. This targeted MLE can now be used to map into a locally efficient estimator of any parameter of μ_0, μ_1 such as the risk difference $\mu_1 - \mu_0$, the relative risk μ_1/μ_0 and the odds ratio $\mu_1(1 - mu_0)/((1-\mu_1)\mu_0)$.

3.5 Estimating the Treatment Mechanism as well

Even when the treatment mechanism (the way treatment was assigned) is known as it is in a randomized trial, it has been shown that efficiency is increased when

estimating it from the data (van der Laan and Robins (2002)). Estimating the treatment mechanism does not add any benefit to the G-computation estimator since it does not use this information. The targeted MLE can however leverage this information to obtain a more precise estimate of the treatment effect. This can be a particular benefit when the model for $Q(A,W)$ is mis-specified. The targeted MLE is still consistent when $Q(A,W)$ is mis-specified, however, we can gain efficiency when estimating the treatment mechanism in such a case. The treatment mechanism can be estimated from the data using a logistic regression model, for example, $\hat{g}^0(1|W) = \frac{1}{1+\exp(-(\alpha_1 W_1 + \alpha_2 W_2))}$, but one can also augment an initial fit $\hat{g}^0$ with a targeted direction aiming for a maximal gain in efficiency: see van der Laan, Rubin (2006). We present the targeted MLE for the risk difference, however, this can be immediately extended to the relative risk and odds ratio as well. Consider the parametric submodel through $\hat{p_0}$ indexed by parameter ϵ,

$$\hat{p}^0(\epsilon)(Y|A,W) = (\hat{Q}^0(\epsilon)(A,W))^Y (1 - \hat{Q}^0(\epsilon)(A,W))^{1-Y}$$

where $\hat{Q}^0(\epsilon)(A,W)$ is given by the logistic regression model,

$$\hat{Q}^0(\epsilon)(A,W) = \frac{1}{1 + \exp -(\hat{m}^0(A,W) + \epsilon h(A,W))}.$$

Setting the score of this model equal to the part of the efficient influence curve that corresponds with scores for $P(Y|A,W)$, and solving for $h(A,W)$ we obtain the covariate,

$$h(A,W) = \frac{I(A=1)}{\hat{g}^0(1|W)} - \frac{I(A=0)}{\hat{g}^0(0|W)},$$

which is added to the logistic regression $\hat{Q}^0(A,W)$. Again,
$\hat{\epsilon} = \arg\max_\epsilon \sum_{i=1}^n \log \hat{Q}^0(\epsilon)(A_i,W_i)$ can be estimated in practice by fitting a logistic regression in $\hat{m}^0(A,W)$ and $h(A,W)$, fixing the coefficient in front of $\hat{m}^0(A,W)$ to 1 and the intercept to 0. The resulting coefficient $\hat{\epsilon}$ for $h(A,W)$ is no longer necessarily equal to 0. Let the targeted MLE for $Q_0(A,W)$ be given by $\hat{Q}^*(A,W) = \hat{Q}^0(\hat{\epsilon})(A,W)$. The targeted MLE for ψ_0 is then,

$$\begin{aligned}
\hat{\psi}_{RD-tMLE2} &= \frac{1}{n}\sum_{i=1}^n \frac{I(A_i=1)}{\hat{g}^0(1|W)}(Y_i - \hat{Q}^*(1,W_i)) - \\
&\quad -\frac{1}{n}\sum_{i=1}^n \frac{I(A_i=0)}{\hat{g}^0(0|W)}(Y_i - \hat{Q}^*(0,W_i)) + \\
&\quad +\frac{1}{n}\sum_{i=1}^n \hat{Q}^*(1,W_i) - \frac{1}{n}\sum_{i=1}^n \hat{Q}^*(0,W_i).
\end{aligned}$$

Note that $\hat{Q}^0(A,W)$ is now updated, contrary to the case when we were not estimating the treatment mechanism as in previous subsections.

3.6 Missing Data

Here we provide the targeted MLE for the case that the outcome Y is subject to missingness that can be informed by the baseline covariates W. In such a case the missingness cannot be ignored as it can lead to biased estimates as treatment groups are no longer balanced with respect to the covariates. Let C represent the indicator whether or not the outcome was observed. The observed data can be represented as $O = (W, A, C, CY) \sim p_0$ and the full data is given by $X = ((Y_a : a \in \mathcal{A}), W)$. We assume that the conditional distribution of the joint censoring variable (A, C) given X satisfies coarsening at random (CAR), i.e. $g_0(A, C|X) = g_0(A, C|W)$. Let

$$P_0 \to \Psi(p_0) = E_{p_0}(P(Y|A = 1, W) - P(Y|A = 0, W))$$

be the parameter of interest. We wish to estimate the risk difference with the targeted MLE. The efficient influence curve is given by,

$$\begin{aligned} D(p_0) \quad = \quad & \frac{I(A=1)}{g_0(1,1|W)}(Y - Q_0(1,1,W)) - \\ & - \frac{I(A=0)}{(g_0(0,1|W))}(Y - Q_0(0,1,W)) + \\ & + Q_0(1,1,W) - Q_0(0,1,W) - \Psi(p_0), \end{aligned}$$

where $g_0(A = 1, c|W) = \delta_0 g(c|A = 1, W)$ and $g_0(A = 0, c|W) = (1 - \delta_0)g(c|A = 0, W)$. We now present the analogue to the derivation of the targeted MLE for ψ_0. Consider the parametric submodel through $\hat{p}^0$ indexed by parameter ϵ,

$$\hat{p}^0(\epsilon)(Y|A, C = 1, W) = (\hat{Q}^0(\epsilon)(A, C = 1, W))^Y (1 - \hat{Q}^0(\epsilon)(A, C = 1, W))^{1-Y}$$

where $\hat{Q}^0(\epsilon)(A, C = 1, W)$ is given by the logistic regression model,

$$\hat{Q}^0(\epsilon)(A, C = 1, W) = \frac{1}{1 + \exp -(\hat{m}^0(A, C = 1, W) + \epsilon h(A, C = 1, W))}.$$

At $C = 0$, the likelihood of $P(Y \mid A, C, W)$ provides as contribution a factor 1, which can thus be ignored. The score for this logistic regression model at $\epsilon = 0$ is given by,

$$\frac{d}{d\epsilon} \log p^0(\epsilon)(A, C, W)\Big|_{\epsilon=0} = I(C = 1)h(A, C = 1, W)(Y - \hat{Q}^0(A, C = 1, W))$$

We now set this score equal to the component of the efficient influence curve which equals a score for $P(Y|A, C = 1, W)$, at $\hat{p}^0$, to obtain the equality

$$\begin{aligned} & h(A, C = 1, W)(Y - \hat{Q}^0(A, C = 1, W)) \\ & \quad = (Y - \hat{Q}^0(A, C = 1, W)) \left(\frac{I(A=1)}{\hat{g}(1,1|W)} - \frac{I(A=0)}{\hat{g}(0,1|W)} \right). \end{aligned}$$

Solving for $h(A, C = 1, W)$ we obtain,

$$h(A, C = 1, W) = \frac{I(A = 1)}{\hat{g}(1, 1|W)} - \frac{I(A = 0)}{\hat{g}(0, 1|W)}.$$

The estimate of ϵ given by $\hat{\epsilon} = \arg\max_\epsilon \sum_{i=1}^n I(C_i = 1) \log \hat{Q}^0(\epsilon)(A_i, W_i)$. Now the logistic regression fit $\hat{Q}^0(Y|A, C = 1, W)$ can be updated by adding as covariate $h(A, C = 1, W)$ to obtain the targeted MLE $\hat{Q}^*(Y|A, C = 1, W)$ for $Q_0(A, C = 1, W)$ based on all observations with $C_i = 1$. The estimate for $P(C = 1|A = 0, W)$ as required to calculate the extra covariate $h(A, W)$ can be obtained by using a logistic regression model selected either data-adaptively or using a fixed pre-specified model for C conditional on $W, A = 0$. The targeted MLE for ψ_0 is given by,

$$\hat{\psi}_{RD-tMLE} = \frac{1}{n} \sum_{i=1}^n \hat{Q}^*(1, 1, W_i) - \hat{Q}^*(0, 1, W_i).$$

We note that the targeted MLE for missing covariate values is derived in exactly the same manner.

4 Testing and Inference

Let $\hat{p}^*$ represent the targeted MLE of p_0. One can construct a Wald-type 0.95-confidence interval based on the estimate of the efficient influence curve, $\hat{IC}(O) = D(\hat{p}^*)$. That is, one can estimate the asymptotic variance of $\sqrt{n}(\hat{\psi} - \psi_0)$ with

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n \hat{IC}^2(O_i).$$

The corresponding asymptotically conservative Wald-type 0.95-confidence interval is defined as $\psi_n \pm 1.96 \frac{\hat{\sigma}}{\sqrt{n}}$. The null hypothesis $H_0 : \psi_0 = 0$ can be tested with the test statistic

$$T_n = \frac{\psi_n}{\frac{\hat{\sigma}}{\sqrt{n}}},$$

whose asymptotic distribution is $N(0, 1)$ under the null hypothesis. We note that this estimate of the asymptotic variance is conservative even if $\hat{Q}^0(A, W)$ is inconsistent, and it is actually asymptotically accurate if $\hat{Q}^0(A, W)$ is consistent (see van der Laan and Rubin (2006) and van der Laan and Robins (2002)). An alternative recommended approach to obtain a non-conservative estimate of the variance is the bootstrap procedure which will provide asymptotically valid confidence intervals.

5 Simulation Studies

5.1 Simulation 1

In this simulation, the treatment A and outcome Y are binary and W is a 2-dimensional covariate, $W = (W_1, W_2)$. The simulated data were generated according to the following laws:

1. $W_1 \sim N(2, 2)$
2. $W_2 \sim U(3, 8)$
3. $P(A = 1) = \delta_0 = 0.5$
4. $Q_0(A, W) = P(Y = 1|A, W) = \frac{1}{(1+\exp(-(kA-5W_1^2+2W_2)))}$

We simulated the data for 2 scenarios based on the value for k in $P(Y = 1|A, W)$. In the first scenario, $k = 1.2$ and there is a small treatment effect and in the second $k = 20$, and there is a larger treatment effect. The risk difference, relative risk and odds ratio were estimated. The true values were given by $P(Y_1 = 1) = 0.372$, $P(Y_0 = 1) = 0.352$ and $(RD, RR, OR) = (0.019, 1.055, 1.087)$ for $k = 1.2$, $P(Y_1 = 1) = 0.583$, $P(Y_0 = 1) = 0.352$ and $(RD, RR, OR) = (0.231, 1.654, 2.570)$ for $k = 20$. The parameters were estimated using 4 methods. The first method "Unadjusted" is the unadjusted method of regressing Y on A using a logistic regression model. The second method "Correct" is the targeted maximum likelihood method which is equivalent to the standard G-computation (maximum likelihood) estimator with $\hat{Q}(A, W) = 1/(1+\exp(-(\hat{\alpha}_0+\hat{\alpha}_1 A+\hat{\alpha}_2 W_1^2+\hat{\alpha}_3 W_2)))$. The third method "Misspec" used a mis-specified fit given by $\hat{Q}(A, W) = 1/(1 + \exp(-(\hat{\alpha}_0 + \hat{\alpha}_1 A + \hat{\alpha}_2 W_1)))$. For the fourth method ,"DSA", the estimate $\hat{Q}(A, W)$ was obtained using Deletion/Substitution/Addition (DSA). The DSA algorithm is a data-adaptive model selection procedure based on cross-validation that relies on deletion, substitution, and addition moves to search through a large space of possible functional forms, and is publicly available at http://www.stat.berkeley.edu/ laan/Software/ (Sinisi and van der Laan (2004)). The variable A was forced into the model and the DSA then selects from the remaining covariates. The maximum power set in the DSA algorithm for any term in the model was set to 2, meaning square terms and 2-way interactions were allowed. Standard errors for the targeted MLE were estimated using the estimated influence curve. For the odds ratio simulations, the estimator obtained by extracting the coefficient for A and the corresponding standard error from the logistic regression model fit is labelled "Adjusted". The simulation was run 1000 times for each sample size: $n = 50, 100, 250, 500, 1000$.

For $k = 1.2$, W strongly predicts Y and thus the targeted MLE, which adjusts for W results in a large increase in efficiency over the unadjusted method as observed by the relative efficiencies (RE) provided in Table 1. The largest gain in efficiency occurs as expected when $\hat{Q}(A, W)$ is correctly specified followed closely by the DSA method, which in general gives a slightly lower bias and

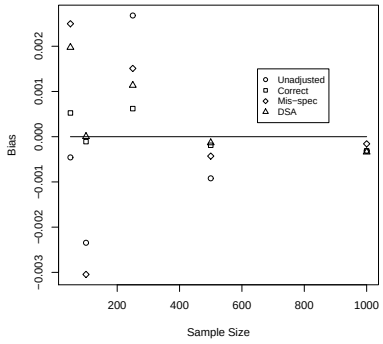
slightly higher variability than the correctly specified model due to overfitting of $\hat{Q}(A, W)$. In the scenario where $k = 20$, A is more strongly predictive of Y as compared to W and thus the increase in efficiency is not as marked as when $k = 1.2$. The largest increase in efficiency for both values of k occurs for the estimates of the odds ratio. When $\hat{Q}(A, W)$ is mis-specified, there is still a noticeable increase in efficiency showing that it is advised to always adjust for covariates. This is a result of the double robustness of the estimator as discussed in section 2. A significant result is the increase in power of the targeted MLE as evidenced by the proportion of rejected tests. In particular when $k = 1.2$, that is when the effect of A is weaker and more difficult to detect, the increase in power is quite significant. When the sample size is greater than 100, and $k = 20$ the unadjusted performs similar to the targeted MLE estimators with respect to power. Another notable result is that the targeted MLE circumvents the issue of singularity, i.e. Y is perfectly predicted by A and W, that occurs when using the adjusted estimate. In this situation the adjusted estimate is drastically inflated and for this reason, the adjusted results were not included in the bias plots. However, this is not an issue for the targeted MLE. The efficiency gain of the targeted MLE increases as the covariate becomes more predictive. This becomes even more drastic when the covariate is perfectly predictive, whereas the adjusted estimate completely breaks down. For example, in a single run of the simulation for the odds ratio with $k = 1.2$, with $n = 50$, the "Adjusted" model fit gave a coefficient of 25.4 and thus an estimate odds ratio of approximately 10^{11}. The corresponding targeted MLE using this same model gives an estimate of 1.083, noting that the true value is 1.087. This is of particular importance for small sample sizes but still occurs even for large sample sizes as shown in the RE estimates for the "Adjusted" estimate in Table 2. We also note that the bias is almost always positive for the relative risk and odds ratios whereas positive and negative bias occurs for the risk difference.

Table 1: Simulation 1: k=1.2: MSE is Mean Squared Error for Unadjusted Estimate, RE is Relative Efficiency of remaining estimators to Unadjusted MSE and Rej is Proportion of Rejected Tests

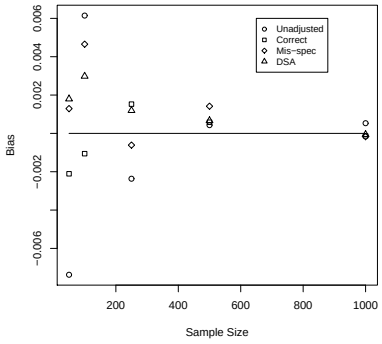
	n=50	n=100	n=250	n=500	n=1000
Risk Difference					
Unadjusted MSE	1.8e-02	9.6e-03	3.5e-03	1.9e-03	8.3e-04
Correct RE	5.41	5.01	10.79	12.25	10.95
Mis-spec RE	2.01	2.31	1.95	2.16	2.10
DSA RE	3.38	7.07	10.72	11.99	10.94
Unadjusted Rej	0.06	0.06	0.06	0.08	0.08
Correct Rej	0.18	0.22	0.27	0.41	0.63
Mis-spec Rej	0.09	0.06	0.09	0.11	0.14
DSA Rej	0.09	0.12	0.27	0.42	0.64
Relative Risk					
Unadj MSE	3.0e-01	1.0e-01	3.6e-02	1.5e-02	7.9e-03
Correct RE	9.08	4.07	12.76	12.55	12.34
Mis-spec RE	2.26	2.36	2.10	2.18	2.06
DSA RE	4.09	7.11	12.03	12.22	12.31
Unadjusted Rej	0.04	0.04	0.06	0.06	0.08
Correct Rej	0.10	0.15	0.22	0.37	0.65
Mis-spec Rej	0.05	0.04	0.06	0.07	0.14
DSA Rej	0.03	0.09	0.22	0.37	0.65
Odds Ratio					
Unadj MSE	1.5e+00	3.1e-01	9.5e-02	4.1e-02	2.0e-02
Adjusted RE	9.4e-178	4.8e-251	5.2e-01	5.3e-01	4.2e-01
Correct RE	1.92	0.00	13.49	13.13	12.78
Mis-spec RE	2.97	2.42	2.39	2.28	1.96
DSA RE	7.07	7.05	13.3	12.72	12.57
Unadjusted Rej	0.04	0.06	0.06	0.06	0.09
Adjusted Rej	0.02	0.04	0.04	0.05	0.13
Correct Rej	0.11	0.14	0.21	0.36	0.67
Mis-spec Rej	0.06	0.04	0.04	0.05	0.14
DSA Rej	0.04	0.07	0.21	0.38	0.68

Table 2: Simulation 1: k=20

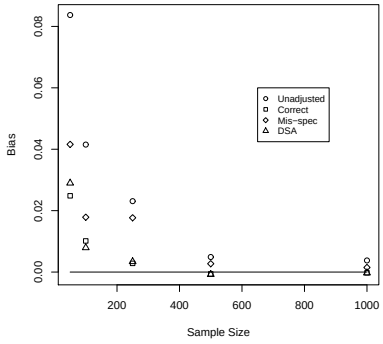
	n=50	n=100	n=250	n=500	n=1000
Risk Difference					
Unadjusted MSE	2.0e-02	9.2e-03	3.9e-03	1.8e-03	9.9e-04
Correct RE	3.80	3.36	4.16	4.22	4.52
Mis-spec RE	2.25	2.45	2.59	2.49	2.50
DSA RE	2.89	3.86	4.33	4.23	4.52
Unadjusted Rej	0.38	0.68	0.95	1.00	1.00
Correct Rej	0.99	1.00	1.00	1.00	1.00
Mis-spec Rej	0.81	0.97	1.00	1.00	1.00
DSA Rej	0.92	1.00	1.00	1.00	1.00
Relative Risk					
Unadj MSE	5.8e-01	2.0e-01	5.5e-02	2.7e-02	1.4e-02
Correct RE	4.76	4.24	3.63	3.98	4.10
Mis-spec RE	2.01	2.22	2.11	2.11	2.19
DSA RE	2.36	3.34	3.34	3.97	4.09
Unadjusted Rej	0.30	0.61	0.94	1.00	1.00
Correct Rej	0.96	1.00	1.00	1.00	1.00
Mis-spec Rej	0.47	0.92	1.00	1.00	1.00
DSA Rej	0.65	0.98	1.00	1.00	1.00
Odds Ratio					
Unadj MSE	6.9e+00	1.9e+00	6.0e-01	2.4e-01	1.2e-01
Adjusted RE	0.00	0.00	1.7e-17	5.4e-03	4.3e-03
Correct RE	0.00	4.58	2.97	4.87	5.01
Mis-spec RE	2.81	2.79	2.63	2.38	2.58
DSA RE	4.59	4.62	5.27	4.82	5.00
Unadjusted Rej	0.33	0.65	0.96	1.00	1.00
Adjusted Rej	0.44	0.89	1.00	1.00	1.00
Correct Rej	0.94	1.00	1.00	1.00	1.00
Mis-spec Rej	0.25	0.84	1.00	1.00	1.00
DSA Rej	0.52	0.98	1.00	1.00	1.00



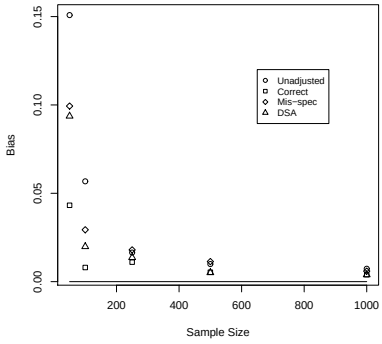
(a) Risk Difference, k=1.2



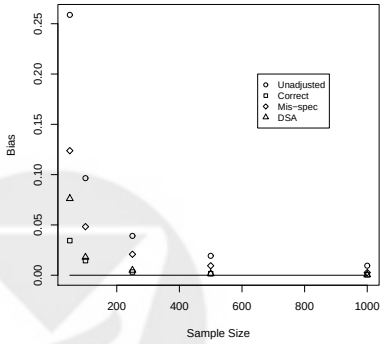
(b) Risk Difference, k=20



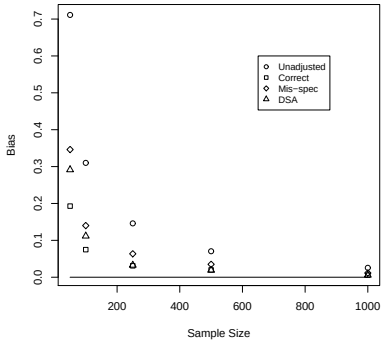
(c) Relative Risk, k=1.2



(d) Relative Risk, k=20



(e) Odds Ratio, k=1.2



(f) Odds Ratio, k=20

Figure 1: Simulation 1: Bias

5.2 Simulation 2: Odds Ratio with Interaction Term

In this simulation, the treatment A and outcome Y are binary and W is a 2-dimensional covariate, $W = (W_1, W_2)$. Here the true causal odds ratio is 0.83. The simulated data were generated according to the following laws:

1. $W_1 \sim N(2, 2)$
2. $W_2 \sim U(3, 8)$
3. $P(A = 1) = \delta_0 = 0.5$
4. $Q_0(A, W) = P(Y = 1|A, W) = \frac{1}{(1+\exp(-(1.2A-5W_1^2+2W_2-5AW_1)))}$

The true values were given by $P(Y_1 = 1) = 0.312$, $P(Y_0 = 1) = 0.352$ and $OR = 0.833$. The same methods used in simulation 1 were used here to estimate the odds ratio. The simulation was run 1000 times for each sample size: $n = 50, 100, 250, 500, 1000$. For the "Mis-spec" targeted MLE, the mis-specified fit was given by $\hat{Q}(A, W) = 1/(1+\exp(-(\hat{\alpha}_0+\hat{\alpha}_1 A+\hat{\alpha}_2 W_1)))$. Figure 2 provides a plot of the bias for each of the estimators. The results are similar to odds ratio for simulation 1 in that the bias is positive for all estimators, and thus the odds ratio is over-estimated. Again, even when $\hat{Q}(A, W)$ is mis-specified the bias and MSE are reduced as compared to the unadjusted estimate (Table 3). The DSA, which allows for interactions, shows a significant improvement in terms of bias and MSE. A notable increase in power is again observed for the targeted MLE over the unadjusted method.

Table 3: Odds Ratio, with Interaction

	50	100	250	500	1000
Unadjusted MSE	5.6e-01	1.6e-01	5.9e-02	2.6e-02	1.2e-02
Adjusted RE	0.00	0.00	0.65	0.56	0.38
Correct RE	7.37	1.67	2.22	7.56	7.71
Mis-spec RE	2.78	2.52	2.44	2.60	2.69
DSA RE	5.30	5.69	6.65	7.26	7.68
Unadjusted Rej	0.05	0.07	0.10	0.17	0.31
Adjusted Rej	0.02	0.05	0.13	0.25	0.50
Correct Rej	0.99	0.99	1.00	1.00	1.00
Mis-spec Rej	0.96	1.00	1.00	1.00	1.00
DSA Rej	0.98	1.00	1.00	1.00	1.00

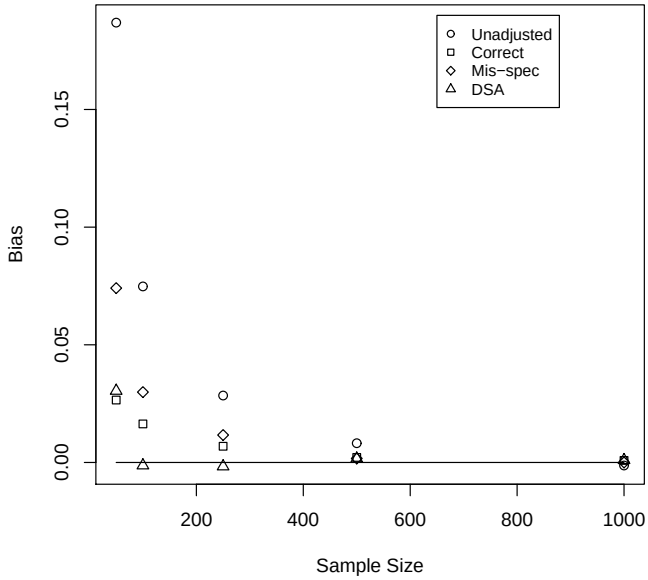


Figure 2: Odds Ratio, Interaction

5.3 Simulation 3: Estimating the Treatment Mechanism as well

In this simulation, the treatment mechanism, $\hat{P}(A|W)$ is estimated from the data using a logistic regression model with covariates that are predictive of the outcome Y. The simulated data were generated according to the following laws:

1. $W_1 \sim N(1, 2)$
2. $W_2 \sim U(1, 4)$
3. $W_3 \sim U(0, 20)$
4. $P(A = 1) = \delta_0 = 0.5$
5. $Q_0(A, W) = P(Y = 1|A, W) = \frac{1}{(1+\exp(-(3A-2W_1^2-\log(W_2)+0.5W_3)))}$

The true values were given by $P(Y_1 = 1) = 0.569$, $P(Y_0 = 1) = 0.419$ and $RD = 0.150$. The treatment mechanism was estimated with the logistic regression model given by $g(A|W) = 1/(1+\exp(-(\gamma_0+\gamma_1 W_1+\gamma_2 W_2+\gamma_3 W_3)))$. The targeted MLE estimator, represented as "Est tx" in Table 5 and Figure

4, with the estimated treatment mechanism is no longer equivalent to the G-computation estimator. The mis-specified fit for $Q(A,W) = 1/(1+\exp(-(\alpha_0 + \alpha_1 A + \alpha_2 W_1)))$ is used as the initial fit and the covariate $h(A,W)$ provided in section 3.4 is then added to this logistic regression. The targeted MLE is then estimated as usual. Thus, we are interested in comparing the mis-specified targeted MLE to the estimated treatment mechanism targeted MLE. Figure 4 shows the bias is reduced and the efficiency is slightly increased when estimating the treatment mechanism. The power was approximately equal for the mis-specified and estimated treatment mechanism targeted MLE. The DSA targeted MLE method again shows a large improvement in efficiency and power over the unadjusted method.

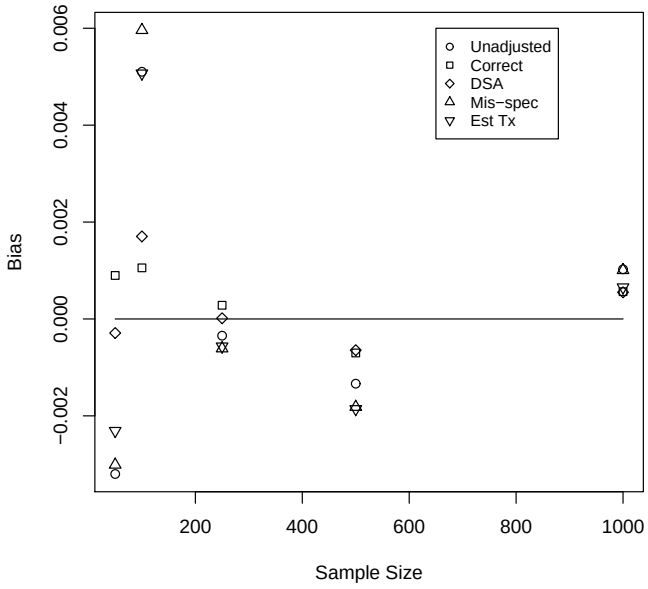


Figure 3: Risk Difference, Estimated Treatment Mechanism

Table 4: Risk Difference, Estimated Tx Mechanism

	50	100	250	500	1000
Unadjusted MSE	2.1e-02	9.4e-03	3.8e-03	1.9e-03	9.9e-04
Correct RE	3.22	3.51	3.91	3.95	4.08
DSA RE	2.65	3.48	3.89	3.94	4.04
Mis-spec RE	1.19	1.18	1.16	1.21	1.20
Est tx RE	1.26	1.30	1.28	1.34	1.29
Unadjusted Rej	0.22	0.34	0.67	0.92	1.00
Correct Rej	0.73	0.90	1.00	1.00	1.00
DSA Rej	0.59	0.90	1.00	1.00	1.00
Mis-spec Rej	0.26	0.42	0.76	0.96	1.00
Est tx Rej	0.23	0.40	0.75	0.96	1.00

5.4 Efficiency Gain and R^2

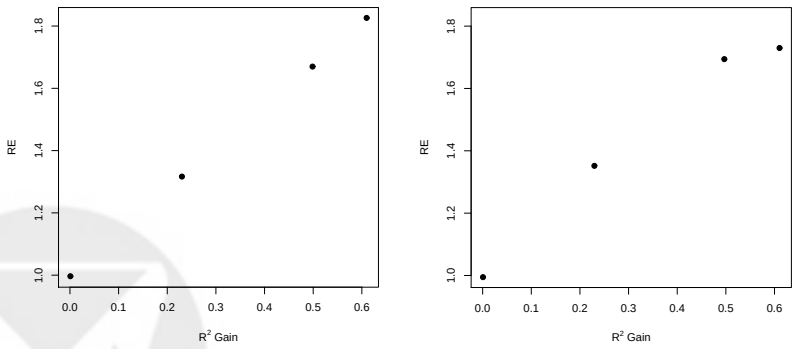
The gain in relative efficiency is related to the gain in the squared multiple correlation coefficient R^2. A covariate predictive of the outcome results in an increase in R^2 in the adjusted model as compared to the unadjusted model. The increase in R^2 results in an increase in efficiency in the targeted MLE. Pocock et al. (2002) discussed the increase in efficiency when adjusting for predictive covariates in linear models. The following simulations show that this also applies to the targeted MLE using logistic regression models. Simulated data were generated according to the following laws:

1. $\sqrt{W} \sim N(2, 2)$
2. $P(A = 1) = \delta_0 = 0.5$
3. $Q_0(A, W) = P(Y = 1|A, W) = \frac{1}{(1+\exp(-(1.2A - cW)))}$

A simulation of sample size $n = 1000$ was run for each $c = \{0, 0.25, 2, 10\}$, that is covariate W is increasingly predictive. The R^2 was estimated in the ordinary least squares sense,

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(Y_i - \hat{Q}(A, W))^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}.$$

A gain in R^2 was computed as the difference between R^2 in the covariate adjusted model and the covariate unadjusted model. Figure 5 and 6 depict the relative efficiency to the unadjusted model for the targeted MLE of the odds ratio against the gain in R^2 for the targeted MLE of the odds ratio and risk difference respectively.



(a) Risk Difference (b) Odds Ratio

Figure 4: Efficiency Gain and R^2

5.5 Simulations Discussion

The 4 simulations were relatively simple scenarios but were useful in demonstrating the following points:

- The targeted MLE shows a clear increase in both efficiency and power over the unadjusted method, even when $Q(A, W)$ is not correctly specified.
- The DSA method for selecting $Q(A, W)$ provides a significant increase in efficiency and power over the mis-specified fixed $Q(A, W)$ method. The average relative efficiencies between these two methods ranged from 1.7 to 3.6 for sample sizes $n = 50$ to $n = 1000$ in our simulations.
- The targeted MLE circumvents the singularity issue that occurs when using the adjusted method of extracting the coefficient from the logistic regression model $Q(A, W)$.
- Interaction terms in the model for $Q(A, W)$ fit entirely into the framework of the targeted MLE.
- Estimating the treatment mechanism provides a further small increase in efficiency over targeting only $Q(A, W)$.

6 Discussion

The targeted MLE provides a general framework that we applied to estimation of the marginal (unadjusted) effect of treatment in randomized trials. We observed that the traditional method of covariate adjustment in randomized trials using logistic regression models can be mapped, by averaging over the covariate(s), to obtain a fully robust and efficient estimator of the marginal effect, which equals the targeted MLE. We demonstrated that the targeted MLE does just this and results in an increase in efficiency and power over the unadjusted method, contrary to what has been reported in the literature for covariate adjustment for logistic regression. The simulation results showed that data-adaptive model selection algorithms such as the DSA, which we used in this paper, or forward selection, when specified a priori should be used. However, we showed that even adjusting by a misspecified regression model results in gain in efficiency and power. Thus, using an a priori specified model, even if it is mis-specified, can increase the power, and thus reduce the sample size requirements for the study. This is particularly important for trials with smaller sample sizes. The targeted MLE framework can also address missing data, either in the outcome as we demonstrated in section 3.5 for the risk difference, but also missingness in covariates and treatment as well for any of the parameters of interest. In these scenarios the targeted MLE covariate may not be as straightforward as those that were presented in this paper, but its derivation is analogue. We focused on logistic and relative risk regression, but the methodology can be extended to any other regression models for $Q(A, W)$. The targeted MLE framework can

also be applied to other parameters of interest in randomized trials such as an adjusted effect, for example by age or biomarker, and can also handle survival times as outcomes (see, van der Laan, Rubin (2006)).

References

S.F. Assmann, S.J. Pocock, L.E. Enos, L.E. Kasten. Subgroup analysis and other (mis)uses of baseline data in clinical trials. *Lancet*, 355:1064 1069, 2000.

H. Bang, J.M. Robins. Doubly Robust Estimation in Missing Data and Causal Inference Models. *Biometrics*, 61:962-972, 2005.

F.J. Belda, L. Aguilera, J.G. de la Asunción, J. Alberti, R. Vicente, L. Ferrándiz, R. Rodríguez, R. Company, D.I. Sessler, G. Aguilar, S.G Botello, R. Ortí, for the Spanish Reduccion de la Tasa de Infeccion Quirurgica Group. Supplemental Perioperative Oxygen and the Risk of Surgical Wound Infection. *Journal of the American Medical Association*, 294:2035-2042, 2005.

N. Frasure-Smith, F. Lespérance, R.H. Prince, P. Verrier, R. Garber, M. Juneau, C. Wolfson, M. Bourassa. Randomised trial of home-based psychological nursing intervention for patients recovering from myocardial infarction. *Lancet*, 350:473479, 1997.

C.M. der Horst, M.S. Saag, G.A. Cloud, R.J. Hamill, J.R. Graybill, J.D. Sobel, P.C. Johnson, C.U. Tuazon, T. Kerkering, B.L. Moskovitz, W.G. Powderly, W.E. Dismukes. The National Institute of Allergy and Infectious Diseases Mycoses Study Group and AIDS Clinical Trials Group. Treatment of cryptococcal meningitis associated with the acquired immunodeficiency syndrome. *New England Journal of Medicine*, 337:1521, 1997.

A.V. Hernández, E.W. Steyerberg and J.D.F. Habbema. Covariate adjustment in randomized controlled trials with dichotomous outcomes increases statistical power and reduces sample size requirements. *Journal of Clinical Epidemiology.* 57(5):454-460, 2004.

R. Neugebauer and M. J. van der Laan. Why prefer double robust estimates? Illustration with causal point treatment studies. Technical Report 115, Division of Biostatistics, University of California, Berkeley, 2002.

S.J. Pocock, S.E. Assmann, L.E. Enos, L.E. Kasten. Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: current practice and problems. *Statistics in Medicine*, 21:29172930, 2002.

A.G. Randolph, D. Wypij, S.T. Venkataraman, J.H. Hanson, R.G. Gedeit, K.L. Meert, P.M. Luckett, P. Forbes, M. Lilley, J. Thompson, I.M. Cheifetz, P. Hibberd, R. Wetzel, P.N. Cox, J.H. Arnold, for the Pediatric Acute Lung Injury and Sepsis Investigators (PALISI) Network. Effect of Mechanical Ventilator Weaning Protocols on Respiratory Outcomes in Infants and Children. *Journal of the American Medical Association*, 288:2561-2568, 2002.

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

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

J.M. Robins. Robust estimation in sequentially ignorable missing data and causal inference models. In *Proceedings of the American Statistical Association*, 2000.

J. M. Robins and A. Rotnitzky. Comment on the Bickel and Kwon article, "Inference for semiparametric models: Some questions and an answer". *Statistica Sinica*, 11(4):920936, 2001.

L.D. Robinson, N.P. Jewell. Some surprising results about covariate adjustment in logistic regression models. *International Statistical Review*, 59:227240, 1991.

D.O. Scharfstein, A. Rotnitzky, and J.M. Robins. Adjusting for nonignorable drop-out using semiparametric nonresponse models. *Journal of the American Statistical Association* 94, 10961120 (with Rejoinder, 1135 1146), 1999.
S. Sinisi and M.J. van der Laan. The deletion/substitution/addition algorithm in loss function based estimation: Applications in genomics. *Journal of Statistical Methods in Molecular Biology*, 3(1), 2004.

A. A. Tsiatis, M. Davidian, M. Zhang, X. Lu. Covariate adjustment for two-sample treatment comparisons in randomized clinical trials: A principled yet flexible approach. Submitted to *Statistics in Medicine* special issue on "Statistical Methods in HIV/AIDS and its practical application." 2006.

M.J. van der Laan and J.M. Robins. Unified methods for censored longitudinal data and causality. Springer, New York, 2002.

M.J. van der Laan and D. Rubin. Targeted Maximum Likelihood Learning. Technical Report 213, Division of Biostatistics University of California, Berkeley, 2006.