The International Journal of Biostatistics

Volume 6, *Issue* 1 2010 *Article* 13

Simple, Efficient Estimators of Treatment Effects in Randomized Trials Using Generalized Linear Models to Leverage Baseline Variables

Michael Rosenblum* Mark J. van der Laan†

*Johns Hopkins University, mrosenbl@jhsph.edu

†University of California, Berkeley, laan@berkeley.edu

Simple, Efficient Estimators of Treatment Effects in Randomized Trials Using Generalized Linear Models to Leverage Baseline Variables

Michael Rosenblum and Mark J. van der Laan

Abstract

Models, such as logistic regression and Poisson regression models, are often used to estimate treatment effects in randomized trials. These models leverage information in variables collected before randomization, in order to obtain more precise estimates of treatment effects. However, there is the danger that model misspecification will lead to bias. We show that certain easy to compute, model-based estimators are asymptotically unbiased even when the working model used is arbitrarily misspecified. Furthermore, these estimators are locally efficient. As a special case of our main result, we consider a simple Poisson working model containing only main terms; in this case, we prove the maximum likelihood estimate of the coefficient corresponding to the treatment variable is an asymptotically unbiased estimator of the marginal log rate ratio, even when the working model is arbitrarily misspecified. This is the log-linear analog of ANCOVA for linear models. Our results demonstrate one application of targeted maximum likelihood estimation.

KEYWORDS: misspecified model, targeted maximum likelihood, generalized linear model, Poisson regression

1 Introduction

The appropriate use of models in analyzing the results of randomized trials has been the focus of many recent papers (e.g. Yang and Tsiatis (2001); Pocock et al. (2002); Rosenbaum (2002); Leon et al. (2003); Tsiatis et al. (2008); Moore and van der Laan (2007); Freedman (2008a,b,c); Zhang et al. (2008); Rosenblum and van der Laan (2009)). We focus on estimating marginal treatment effects, such as the risk difference, risk ratio, and log odds ratio. The model-based estimators we present are asymptotically unbiased, and leverage baseline variables to try to get more precision than estimators that ignore baseline variables. All of our results hold even when the models used are arbitrarily misspecified, that is, when the models used do not contain the true data generating distribution. This is an important property since in practice, models will often be misspecified. Our results demonstrate an application of targeted maximum likelihood estimation, a general estimation method with broad applicability to randomized trials and observational studies described in (van der Laan and Rubin, 2006; Moore and van der Laan, 2007; Polley and van der Laan, 2009; van der Laan et al., 2009).

In the next section, we describe the estimation problem being considered and present related work. Our class of estimators and our main result are presented in Section 3. We note that our estimators are examples of doubly robust estimators (Robins, 2000; Robins and Rotnitzky, 2001; Neugebauer and van der Laan., 2002; van der Laan and Robins, 2002), and that the theory already developed for these estimators implies their robustness to misspecification of the generalized linear model used. The contributions in this paper are (1) to show the implementation of targeted maximum likelihood estimation applied to randomized trial data and (2) to highlight a useful class of easy to compute estimators that leverage baseline variables yet are guaranteed to be consistent estimators of marginal mean effects. We show how to construct confidence intervals and compute p-values in Section 4. We present a simulation study comparing the power of our estimators to the unadjusted estimator in Section 5. In Section 6 we give a brief overview of targeted maximum likelihood methodology, of which our estimators are one application. Proofs of our results are given in the Appendix.

2 Description of Estimation Problem, Assumptions, and Related Work

We consider a randomized trial with n subjects. The data collected on each subject include baseline variables (measured before randomization), the random treatment assignment, and the outcome. We assume subjects are randomized with probability 1/2 to either the treatment or control arm, independent of the baseline variables. We let A denote the treatment assignment, with $A = 1$ corresponding to the treatment arm and $A = 0$ corresponding to the control arm. We denote the outcome variable by Y, which may be continuous or discrete valued. We let V denote a subset of the baseline variables (which must be chosen before the trial starts). For each subject i, we denote their data by the vector (V_i, A_i, Y_i), representing baseline measurements, treatment assignment, and outcome, respectively. In general, we recommend that baseline variables that are highly predictive of the outcome should be included in the vector V.

2.1 Parameter We Will Estimate

We consider estimation and inference for parameters that are smooth functions r of the mean effects of being assigned to the two study arms: $E(Y|A = 0)$ and $E(Y|A = 1)$. This class of parameters includes the difference in means $E(Y|A = 1) - E(Y|A = 0)$ (corresponding to $r(x, y) = y - x$), the ratio of means (or rate ratio) $E(Y|A = 1)/E(Y|A = 0)$ (corresponding to $r(x, y) = y/x$), and the log odds ratio $\log \frac{P(Y=1|A=1)/(1-P(Y=1|A=1))}{P(Y=1|A=0)/(1-P(Y=1|A=0))}$ (corresponding to $r(x, y) = \log y(1 - x)/(x(1 - y))$), for example. (Throughout the paper "log" refers to the natural logarithm.) The "unadjusted estimator" estimates these parameters by substituting the sample means in the control arm and treatment arm, respectively, for $E(Y|A = 0)$ and $E(Y|A = 1)$. We will denote $E(Y|A = 0)$ and $E(Y|A = 1)$ by E_0 and E_1, respectively. We are estimating marginal effects of treatment, that is, the overall average effect on a population, comparing two treatments. Such comparisons play a role, for example, when the U.S. Food and Drug Administration (FDA) makes decisions whether to approve new drugs. However, we note that in some problems, for example in some cases when considering repeated measurements (Lindsey and Lambert, 1998), it may be of more interest to estimate conditional effects.

DOI: 10.2202/1557-4679.1138

2.2 Assumptions on the Data Generating Distribution

We assume that each vector of observations (V_i, A_i, Y_i) is an independent, identically distributed draw from an unknown data generating distribution $p^*(V, A, Y)$.[1] We also assume the values of all variables are bounded. We assume that A and V are independent, which is ensured by randomization. These are the only assumptions we make about the data generating distribution (except for the assumptions (i) and (ii) given in Section 2.3 below, both of which can be verified from the data, with probability tending to 1 as sample size goes to infinity). The estimators we give below can be extended to the case where treatment assignment A depends on V, as we discuss briefly in Section 7.

2.3 Requirements on the Form of the Working Model

We will use the machinery of maximum likelihood estimation for generalized linear models, but will not assume the models used are in any way correctly specified. The generalized linear model family (e.g. Binomial, Normal, Poisson), the link function, and the terms in the linear part of the model can all be incorrect. We furthermore allow that the true data generating distribution may not be contained in any generalized linear model at all. The only assumptions we make on the data generating distribution are those listed in Section 2.2 above; we will show in Theorem 1 in Section 3 that regardless of the type of misspecification, our estimators are consistent (i.e. asymptotically unbiased) and asymptotically normal (so asymptotically correct confidence intervals can be obtained). We consider the generalized linear models merely as working models, that is, formulas that are input along with data into an algorithm (such as the targeted maximum likelihood algorithm). When the working models are correctly specified, we will obtain estimators that are optimal in terms of asymptotic mean squared error[2]. When the working models are misspecified, we are still guaranteed that our estimators are consistent (i.e. converge to the true value of the parameter) and asymptotically normal.

Before listing the requirements that our working models must satisfy, we give some examples of generalized linear models that meet these requirements. We note that there is much flexibility in choosing which terms to use in the

[1] This assumption is not guaranteed by randomization. For discussion of this issue, see (Rosenbaum, 2002; Freedman, 2008c; Rosenblum and van der Laan, 2009).

[2] More precisely, when the working models used are correctly specified, our estimators will achieve the semiparametric efficiency bound, and so have minimal asymptotic variance among all regular, asymptotically linear estimators.

linear part of the model; for example, one can include multiple baseline variables, any set of interactions, and any functions of baseline variables and the treatment, as long as the restrictions given below are adhered to. We give just a few, simple possibilities below.

Examples of Working Models Satisfying Requirements Above:

- Least Squares Regression: For Y continuous, the Normal model where $E(Y|A,V)$ is modeled by:

  $$\mu_1(A,V|\beta) = \beta_0 + \beta_1 A + \beta_2 V + \beta_3 AV,$$

- Logistic Regression: For Y binary and $\text{logit}(x) = \log(x/(1-x))$, the following model for $P(Y=1|A,V)$:

  $$\mu_2(A,V|\beta) = \text{ logit}^{-1}\left(\beta_0 + \beta_1 A + \beta_2 V\right),$$

- Poisson Regression: For Y a "count" (that is, Y a nonnegative integer), the Poisson (log-linear) model with mean of Y given A,V of the form:

  $$\mu_3(A,V|\beta) = \exp\left(\beta_0 + \beta_1 A + \beta_2 V\right).$$

- Gamma Regression: For Y positive, real valued, the Gamma model with mean of Y given A,V modeled by:

  $$\mu_4(A,V|\beta) = 1/\left(\beta_0 + \beta_1 A + \beta_2 \exp(V) + \beta_3 \exp(AV)\right),$$

  where each coefficient β_i is restricted to be nonnegative and β_0 is restricted to be bounded away from 0 by some $\delta > 0$.

- Inverse Normal Regression: For Y positive, real valued, the Inverse Normal model with mean of Y given A,V modeled by:

  $$\mu_5(A,V|\beta) = 1/\sqrt{\beta_0 + \beta_1 A + \beta_2 \exp(V)},$$

  where each coefficient β_i is restricted to be nonnegative and β_0 is restricted to be bounded away from 0 by some $\delta > 0$.

In the above Gamma regression model and Inverse Normal regression model, we chose parameterizations that ensure the mean will be bounded. This is guaranteed by restricting all components β_i to be nonnegative, by restricting β_0 to be greater than some positive constant, and by making sure each term in the linear part is a nonnegative function of A,V.

We require that the generalized linear models used as working models have canonical link functions, and are from the commonly used families: Normal, Binomial, Poisson, Gamma, or Inverse Normal (see McCullagh and Nelder

DOI: 10.2202/1557-4679.1138

(1998) for definitions of these exponential families). As described in (McCullagh and Nelder, 1998), the density of the outcome Y, conditional on A, V, under such a generalized linear model can be represented, for suitable choices of functions b, c as

$$\exp(Y\eta - b(\eta) + c(Y, \phi)), \tag{1}$$

where $\eta = \sum_j \beta_j f_j(A, V)$ is the linear part of the model, with terms $f_j(A, V)$ and coefficients β_j, and ϕ is a dispersion parameter.[3] We require that the first two terms in the linear part consist of an intercept and the treatment variable A. We require the functions f_j be chosen to be bounded on compact subsets of $\{0, 1\} \times \mathbf{R}^d$, where V is a d-dimensional vector of baseline variables. The canonical link function g is defined as $\dot{b}^{-1}$, the inverse of the derivative of the function b. We let $\mu(A, V)$ denote the mean of Y given A, V according to the density (1), where the dependence of $\mu(A, V)$ on β is implicit. We note that $\mu(A, V) = \dot{b}(\eta(A, V))$, which is proved in (Bickel and Doksum, 2001).

We make two further assumptions, given below, that involve both the data generating distribution and the form of the working model. These are standard regularity conditions required to guarantee convergence of parameter estimates of (possibly misspecified) generalized linear models to some limit value, as sample size goes to infinity.

(i) We assume the terms $f_j(A, V)$ are linearly independent. This means that if for a set of constants c_j we have $\sum_j c_j f_j(A, V) = 0$ a.s., then $c_j = 0$ for all j.

(ii) We assume that there exists a maximizer β^* of the expected log-likelihood

$$E_{p*}\left[Y\eta - b(\eta) + c(Y, \phi)\right] = E_{p*}\left[Y \sum_j \beta_j f_j(A, V) - b\left(\sum_j \beta_j f_j(A, V)\right) + c(Y, \phi)\right], \tag{2}$$

where the expectation is with respect to the true (but unknown) data generating distribution $p^*(V, A, Y)$. We also assume each component of β^* has absolute value smaller than some pre-specified bound M.

One can detect whether (i) or (ii) are violated, based on the data, with probability tending to 1 as sample size tends to infinity. This follows for (i)

[3]For binary outcomes, the function $b(\eta) = \log(1 + e^\eta)$ and $c(Y, \phi) = 0$. For Poisson regression, in which the outcome is a nonnegative integer, $b(\eta) = e^\eta$ and $c(Y, \phi) = -\log Y!$. Note that in both cases, $\ddot{b}(\eta) := \frac{d^2 b}{d\eta^2} > 0$ for all η.

because linear dependencies in the terms in the model will be detected by standard statistical software for sample size larger than the number of terms. Violation of assumption (ii) can be detected, for large enough sample size n (with probability converging to 1), as described in the Appendix.[4]

For the Gamma and Inverse Normal families, where the outcome variable is assumed to take values in $(0, \infty)$, we additionally require that f_j take nonnegative values; also for these two families we restrict β_j to take nonnegative values and that the intercept β_0 be bounded away from 0 by some $\delta > 0$. These requirements are needed due to the form of the canonical link functions for the Gamma and Inverse Normal families ($1/\mu$ and $1/\mu^2$, respectively), which may be unbounded unless restrictions are imposed on the linear part of the model. Furthermore, for these families, we assume that there exists a maximizer β^* of the expected log-likelihood for which all components of β^* are nonnegative and for which β_0^* is strictly greater than δ; just as for (i) and (ii) above, one can detect whether this assumption is violated, based on the data, with probability tending to 1 as sample size tends to infinity.

A key to our results is that in a randomized trial (that is, where by design the treatment A is independent of the baseline variables V), we can write

$$E(Y|A=0) = E_V[E(Y|A=0,V)],$$
$$E(Y|A=1) = E_V[E(Y|A=1,V)].$$

Our strategy is to use a generalized linear model as a working model to estimate $E(Y|A,V)$; in particular, we'll use generalized linear models with canonical link functions and that include an intercept term and a treatment term A. After fitting these models, we evaluate the right hand side of the previous display using the model fit in place of $E(Y|A=0,V), E(Y|A=1,V)$ and replacing the expectation E_V by expectation with respect to the empirical distribution of V. This allows us to leverage baseline variables and potentially improve the power of our estimators compared to unadjusted estimators. We show in Section 3 and the Appendix that this estimation strategy is an example of a targeted maximum likelihood estimator.

The results we prove in Section 3 are asymptotic in the sample size (that is, we prove our estimators are consistent and have a normal distribution in the limit as sample size tends to infinity.) We assume the working model is fixed, that is, it does not change with sample size and is not data-dependent. This can be assured, for example, if the working model is chosen prior to looking at

[4]The argument that one can detect whether (ii) is violated, for large sample size, relies on the concavity of the expected log-likelihood in arbitrarily misspecified generalized linear models with canonical link; this is described in the Appendix.

DOI: 10.2202/1557-4679.1138

the data. However, we discuss in Section 8 how one can generalize our main result to incorporate certain data-adaptive model selection procedures.

2.4 Related Work

Moore and van der Laan (2007) applied targeted maximum likelihood methodology to prove that for randomized trials, certain easy to compute estimators based on a logistic regression working model are asymptotically unbiased (and locally efficient) even when the working model is misspecified. Our results generalize this important result to a larger class of generalized linear models that includes Normal (Gaussian) models, Poisson models with log link, and models based on the Gamma distribution (with reciprocal link) and Inverse Gaussian distribution (with link $1/\mu^2$). We note that estimation of the risk difference using a Normal working model with only main terms corresponds to ANCOVA (analysis of covariance), which has been shown to be asymptotically unbiased even when the model is misspecified (Yang and Tsiatis, 2001; Leon et al., 2003); similar results have been shown for an estimator called ANCOVA II (Yang and Tsiatis, 2001; Leon et al., 2003), which involves ordinary least squares regression of the centered outcome on centered main terms and an interaction term. We also note that in the special case of logistic regression, Freedman (2008c) proved a related result under the framework of randomization inference.

Our result for the special case of a Poisson model with only main terms (see the Corollary in Section 3) is a generalization of a result of Gail (1986) that required much stronger assumptions than used here.

Robinson and Jewell (1991) compare the precision of estimators of the marginal effect and estimators of the conditional effect of a treatment, based on linear and logistic regression models. In this paper we focus only on estimating marginal effects, that is, comparisons of $E(Y|A=1)$ and $E(Y|A=0)$. These are the same quantities estimated by the unadjusted estimator defined above. Our estimators leverage baseline variables in order to try to get more precise (i.e. smaller asymptotic variance) estimates than the unadjusted estimator. We note that whether marginal effects or conditional effects are more relevant will depend on the application at hand. Though the focus of this paper is estimation and inference, certain results have been shown for hypothesis testing, in which certain model-based tests have asymptotically correct Type I error even when working models are arbitrarily misspecified (Rosenblum and van der Laan, 2009).

The estimators we present in Section 3, which are examples of targeted maximum likelihood estimators, also coincide with certain g-computation es-

timators (Robins, 1986, 1987), doubly-robust estimators (Robins, 2000; Robins and Rotnitzky, 2001; Neugebauer and van der Laan., 2002; van der Laan and Robins, 2002), and estimators in (Tsiatis, 2006; Zhang et al., 2008). We note that in general, targeted maximum likelihood estimators will differ from from g-computation estimators, doubly-robust estimators, and estimators in (Tsiatis, 2006; Zhang et al., 2008). In general, some advantages of targeted maximum likelihood estimators include being substitution estimators (so that global constraints, such as estimates being in the range $[0, 1]$ when estimating a probability, are satisfied), and not suffering from multiple solutions as some estimating functions do. For a full list of such advantages, see (van der Laan, 2010).

3 Main Result

Below we present our class of simple estimators based on generalized linear models that are asymptotically unbiased even when the working model used is incorrectly specified. We then give the main result of the paper in Theorem 1. We illustrate the theorem with two examples based on Poisson regression models.

The class of estimators is constructed as follows, for any fixed, generalized linear model with canonical link function, and any continuously differentiable function r:

1. Estimate the coefficients $\{\beta_j\}$ in the linear part of the generalized linear model using maximum likelihood estimation.[5]

2. Compute $\hat{E}_0 := \frac{1}{n}\sum_{i=1}^{n} \hat{\mu}(0, V_i)$, and $\hat{E}_1 := \frac{1}{n}\sum_{i=1}^{n} \hat{\mu}(1, V_i)$, where $\hat{\mu}(a, v)$ is the estimated mean of Y given $A = a, V = v$, based on the fit of the generalized linear model, for treatment assignment a and baseline variables v. $\hat{\mu}$ is formally defined in the Appendix, where we also give R code for computing $\hat{E}_0$ and $\hat{E}_1$.

3. Compute $r(\hat{E}_0, \hat{E}_1)$; this is our estimator of the parameter $r(E(Y|A = 0), E(Y|A = 1))$.

4. Confidence intervals can be obtained based on estimates of the efficient influence function (as described in Section 4) or based on the nonparametric bootstrap.

[5]If the terms in the linear part are linearly dependent, or if the maximum likelihood algorithm fails to converge to a finite value, we consider the estimator to be undefined.

DOI: 10.2202/1557-4679.1138

We have the following theorem stating that the above estimator is asymptotically unbiased and locally efficient.

Theorem 1: *Consider any generalized linear model from the Normal, Binomial, Poisson, Gamma, or Inverse Gaussian family, with canonical link function, in which the linear part contains the treatment variable as a main term and also contains an intercept. Let r be any continuously differentiable function. Under the assumptions in Section 2, the above procedure gives an asymptotically unbiased estimator for the parameter $r(E(Y|A=0), E(Y|A=1))$, even when the generalized linear model is misspecified. The confidence intervals constructed in Section 4 have asymptotically correct coverage, even when the model is misspecified. Furthermore, this estimator is locally efficient in that when the generalized linear model is correctly specified this estimator attains the efficiency bound for the model that only assumes treatment assignment A is independent of baseline variables V.*

The class of estimators in Theorem 1 is derived from targeted maximum likelihood methodology (van der Laan and Rubin, 2006), as described in the Appendix. We point out that in this special case of a randomized trial (the case considered throughout this paper), the particular version of the targeted maximum likelihood estimator given in the Appendix coincides with certain g-computation estimators (Robins, 1986, 1987), doubly-robust estimators (Robins, 2000; Robins and Rotnitzky, 2001; Neugebauer and van der Laan., 2002; van der Laan and Robins, 2002), and estimators in (Tsiatis, 2006; Zhang et al., 2008). In addition, the estimator given in Theorem 1 solves the doubly robust estimating equation, and thereby the theory of statistical inference developed in (van der Laan and Robins, 2002) applies. In general, targeted maximum likelihood estimators will differ from from g-computation estimators, doubly-robust estimators, and estimators in (Tsiatis, 2006; Zhang et al., 2008).

3.1 Application of Main Result in Several Examples

To illustrate the above theorem, consider a Poisson working model with log link function, and linear part $\eta = \beta_0 + \beta_1 A + \beta_2 V$. We will estimate the marginal log rate ratio of the treatment compared to the control: $\log(E(Y|A=1)/E(Y|A=0))$, using this Poisson model as working model (but not assuming it is correctly specified). This corresponds to choosing the function r in the theorem to be $r(x,y) = \log(y/x)$. We follow the steps given above the theorem to compute an estimate of the marginal log rate ratio. First, we use maximum likelihood estimation to produce estimates $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$ for the coefficients $\beta_0, \beta_1, \beta_2$. Next,

we compute

$$\hat{E}_0 := \frac{1}{n}\sum_{i=1}^{n}\hat{\mu}(0, V_i) = \frac{1}{n}\sum_{i=1}^{n}\exp(\hat{\beta}_0 + \hat{\beta}_2 V_i)$$

and

$$\hat{E}_1 := \frac{1}{n}\sum_{i=1}^{n}\hat{\mu}(1, V_i) = \frac{1}{n}\sum_{i=1}^{n}\exp(\hat{\beta}_0 + \hat{\beta}_1 + \hat{\beta}_2 V_i).$$

Lastly, we compute

$$r(\hat{E}_0, \hat{E}_1) = \log(\hat{E}_1/\hat{E}_0) = \log[\sum_{i=1}^{n}\exp(\hat{\beta}_0 + \hat{\beta}_1 + \hat{\beta}_2 V_i)/\sum_{i=1}^{n}\exp(\hat{\beta}_0 + \hat{\beta}_2 V_i)].$$

In this special case, we see that the above estimator can be simplified, leaving as final estimate $\hat{\beta}_1$, the coefficient of the treatment term. Thus, the above theorem implies the following corollary (which is the log-linear analog of ANCOVA for linear models):

Corollary: *Consider a Poisson working model with only main terms A and V (where V is a vector of pre-randomization variables). Under the assumptions in Section 2, we have $\hat{\beta}_1$, the estimate of the coefficient corresponding to the treatment term A, is an asymptotically unbiased estimate of the marginal log rate ratio, even when the model is misspecified. Also, the confidence intervals constructed in Section 4 have asymptotically correct coverage. Furthermore, this estimator is locally efficient in that when the Poisson model is correctly specified this estimator attains the efficiency bound for the model that only assumes treatment assignment A is independent of baseline variables V.*

As another example, consider the problem of estimating the marginal log rate ratio using a Poisson working model with log link function, but this time with the linear part containing an interaction term: $\eta = \beta_0 + \beta_1 A + \beta_2 V + \beta_3 AV$. Such a working model might be used when it is suspected that there is effect modification by the baseline variable V; in this case, it still may be of interest to estimate the marginal log rate ratio for the total population, recognizing that the parameters E_0, E_1 each represent aggregated mean outcomes over the total population, for the control and treatment interventions, respectively. Again, we use maximum likelihood estimation to get estimates for the coefficients

DOI: 10.2202/1557-4679.1138

$\beta_0, \beta_1, \beta_2, \beta_3$; as above, we use as estimator $r(\hat{E}_0, \hat{E}_1)$, which equals

$$\begin{aligned} \log(\hat{E}_1/\hat{E}_0) &= \log\left[\sum_{i=1}^{n} \exp(\hat{\beta}_0 + \hat{\beta}_1 + \hat{\beta}_2 V_i + \hat{\beta}_3 V_i) / \sum_{i=1}^{n} \exp(\hat{\beta}_0 + \hat{\beta}_2 V_i)\right] \\ &= \hat{\beta}_1 + \log\left[\sum_{i=1}^{n} \exp((\hat{\beta}_2 + \hat{\beta}_3) V_i) / \sum_{i=1}^{n} \exp(\hat{\beta}_2 V_i)\right]. \end{aligned} \quad (3)$$

The proof of Theorem 1, given in the Appendix, applies the targeted maximum likelihood algorithm to the application in this paper, namely, estimating a function r of the conditional means given assignment to the control arm and the treatment arm, respectively. It turns out in this case, that when the initial density p_0 is chosen based on the maximum likelihood estimate using a generalized linear model for Y given A, V, and a canonical link is used, then the targeted maximum likelihood algorithm converges in a single iteration (as defined in Section 6 below) and produces the estimator $r(\hat{E}_0, \hat{E}_1)$ given in steps 1-4 just before Theorem 1. The reason is that the score of a generalized linear model with canonical link function has a simple form that is closely related to the efficient influence function of the conditional means $E(Y|A = 0)$ and $E(Y|A = 1)$ in the model that is nonparametric except for assuming A is randomized (that is, independent of baseline variables V).

4 Computing Confidence Intervals and p-values

We show how to compute confidence intervals and p-values for the estimator given just before Theorem 1. We use the method from Section 4 of (Moore and van der Laan, 2007), based on estimates of the efficient influence function of our parameter in the nonparametric model. This involves first computing an estimate $\hat{\sigma}^2$ for the asymptotic variance of $\sqrt{n}(\hat{\psi} - \psi)$, where $\hat{\psi} = r(\hat{E}_0, \hat{E}_1)$ is our estimator and ψ is the true (but unknown) value for the parameter $r(E_0, E_1)$ that we are estimating; we describe how to compute $\hat{\sigma}^2$ below. Having computed $\hat{\sigma}^2$, we next compute a 95% confidence interval $(\hat{\psi} - 1.96\hat{\sigma}/\sqrt{n}, \hat{\psi} + 1.96\hat{\sigma}/\sqrt{n})$. Also, we can test the null hypothesis $\psi = \psi_0$ using the test statistic $T = \sqrt{n}(\hat{\psi} - \psi_0)/\hat{\sigma}$, which is asymptotically normally distributed with mean 0 and variance 1 under this null hypothesis and under the regularity conditions given in Section 2. The confidence interval and p-value computed by this method are asymptotically correct, even when the generalized linear model used is incorrectly specified. We note that an alternative method to what we present in this section is to use the nonparametric bootstrap.

The above procedures rely on an estimate of the asymptotic variance of $\sqrt{n}(\hat{\psi}-\psi)$, which we denote by $\hat{\sigma}^2$, and define now. It can be computed based on the partial derivatives of the function r used in defining our parameter and on estimates of the efficient influence function of (E_0, E_1) in the nonparametric model. Let r_1', r_2' denote the partial derivatives of the function r with respect to the first component and second component, respectively. For example, when our parameter is the marginal log rate ratio, then $r(x,y) = \log y/x$, and so $r_1'(E_0,E_1) = -1/E_0, r_2'(E_0,E_1) = 1/E_1$. Define the vector with two components $D(p)(V,A,Y) := (D_1(p)(V,A,Y), D_2(p)(V,A,Y))$, where:

$$D_1(p)(V,A,Y) := \tfrac{(1-A)(Y-E_p(Y|A=0,V))}{p(A=0)} + E_p(Y|A=0,V) - E_p(Y|A=0), \quad (4)$$

$$D_2(p)(V,A,Y) := \tfrac{A(Y-E_p(Y|A=1,V))}{p(A=1)} + E_p(Y|A=1,V) - E_p(Y|A=1), \quad (5)$$

where E_p is the expectation with respect to the density p. $D(p)$ is the efficient influence function for (E_0, E_1) at p in the nonparametric model. (See van der Laan and Robins (2002) for the derivation of this efficient influence function.)

As in Theorem 1, assume there exists a maximizer β^* of the expected log-likelihood of the generalized linear model, where the expectation is with respect to the true (but unknown) data generating distribution. Under this assumption, we show in the Appendix that such a maximizer is unique, and that the maximum likelihood estimator $\hat{\beta}$ converges to β^*. Let $p_1(\beta^*)$ be the density of Y given A, V corresponding to the parameter $\beta = \beta^*$ in the generalized linear model. Let p_2 be the known density of A given V, and let p_3 denote the (unknown) marginal density of V. Let density p^* be that of the true (but unknown) data generating distribution. In terms of D and r_1', r_2', the asymptotic variance of $\sqrt{n}(\hat{\psi}-\psi)$ is

$$\begin{aligned}\sigma^2 &= E_{p^*}\left[r_1'(E_0,E_1)D_1(p_1(\beta^*)p_2p_3)(V,A,Y)\right. \\ &\qquad \left. + r_2'(E_0,E_1)D_2(p_1(\beta^*)p_2p_3)(V,A,Y)\right]^2. \end{aligned} \quad (6)$$

We estimate this by

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(r_1'(\hat{E}_0,\hat{E}_1)D_1(\hat{p})(V_i,A_i,Y_i) + r_2'(\hat{E}_0,\hat{E}_1)D_2(\hat{p})(V_i,A_i,Y_i)\right)^2, \quad (7)$$

where $\hat{p}$ is the density estimated by targeted maximum likelihood given in the Appendix. Since as shown in the Appendix, $E_{\hat{p}}(Y|A=0,V) = \hat{\mu}(0,V)$ and $E_{\hat{p}}(Y|A=1,V) = \hat{\mu}(1,V)$, where $\hat{\mu}(a,v)$ is the predicted mean of Y given $A=a, V=v$ based on the maximum likelihood estimate for the generalized

DOI: 10.2202/1557-4679.1138

linear model, we have

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\Big(r_1'(\hat{E}_0,\hat{E}_1)[(1-A_i)(Y_i-\hat{\mu}(0,V_i))/(1/2)+\hat{\mu}(0,V_i)-\hat{E}_0] + r_2'(\hat{E}_0,\hat{E}_1)[A_i(Y_i-\hat{\mu}(1,V_i))/(1/2)+\hat{\mu}(1,V_i)-\hat{E}_1]\Big)^2.$$

For example, when our parameter is the marginal log rate ratio, so that as argued above $r_1'(E_0,E_1)=-1/E_0$, $r_2'(E_0,E_1)=1/E_1$, we have

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(-\frac{1}{\hat{E}_0}\left[(1-A_i)(Y_i-\hat{\mu}(0,V_i))/(1/2)+\hat{\mu}(0,V_i)-\hat{E}_0\right] + \frac{1}{\hat{E}_1}\left[A_i(Y_i-\hat{\mu}(1,V_i))/(1/2)+\hat{\mu}(1,V_i)-\hat{E}_1\right]\right)^2. \quad (8)$$

Having now computed $\hat{\sigma}^2$, one can use this in the formulas given in the first paragraph of this section to compute confidence intervals and p-values.

We note that the above procedure for constructing confidence intervals is invariant to affine transformations of the parameter $r(E_0,E_1)$, but is not invariant to more general monotone transformations of $r(E_0,E_1)$. For a given, strictly monotone function h, one can apply the method of this section to the function $h\circ r$, to obtain a confidence interval for the parameter $h(r(E_0,E_1))$. An alternative method for constructing a confidence interval for $h(r(E_0,E_1))$ is to first compute a confidence interval $[l,u]$ for $r(E_0,E_1)$ as above, and then let $[h(l),h(u)]$ be one's confidence interval for $h(r(E_0,E_1))$. Though these two methods yield confidence intervals that are asymptotically equivalent, they may have different performance for finite samples.

5 Simulation Study

We show the finite sample performance of our proposed estimators for several simulated data generating distributions. We focus on nonnegative integer-valued outcome variables Y. Our parameter in all the simulations is the marginal log rate ratio. We will compare the targeted maximum likelihood estimator to the unadjusted estimator in terms of mean squared error, relative efficiency, and coverage probability of the associated 95% confidence intervals.

5.1 Estimators

The targeted maximum likelihood estimator will use the log-linear working model with main terms and an interaction term, as in the last example given

in Section 3. That is, the working model for Y given A, V is Poisson with linear part containing the terms $1, A, V, AV$, and using canonical link (the log link). The resulting estimator is that given by (3).

The unadjusted estimator is the log of the ratio of the sample means $\hat{\mu}_1, \hat{\mu}_0$ in the treatment and control groups, respectively.

Both methods compute confidence intervals using the delta method, as in Section 4. That is, for the unadjusted estimator $\hat{\psi}_{\text{unadj}}$, we get an estimate for the asymptotic variance of $\sqrt{n}(\hat{\psi}_{\text{unadj}} - \psi)$ using the formula

$$\hat{\sigma}^2_{\text{unadj}} := \frac{1}{n}\sum_{i=1}^{n}\left(-\frac{2}{\hat{\mu}_0}\left[(1-A_i)(Y_i-\hat{\mu}_0)\right] + \frac{2}{\hat{\mu}_1}\left[A_i(Y_i-\hat{\mu}_1)\right]\right)^2. \qquad (9)$$

We use the formula (8) to obtain the analogous asymptotic variance corresponding to the targeted maximum likelihood estimator. Given estimator $\hat{\psi}$ and estimated asymptotic variance $\hat{\sigma}^2$, we form 95% confidence intervals: $(\hat{\psi} - 1.96\hat{\sigma}/\sqrt{n}, \hat{\psi} + 1.96\hat{\sigma}/\sqrt{n})$.

5.2 Data Generating Distributions

We consider three data generating distributions. In all of them, the baseline variable V is a standard normal, and A is binary, independent of V, and takes values 0 and 1 each with probability 1/2.

Under the first data generating distribution (to be defined next), the working model will be correctly specified; under the latter data generating distributions, the working model will be misspecified.

In data generating distribution 1, we set Y to have a Poisson distribution, with $E(Y|A,V) := \exp(A + AV)$. In data generating distribution 2, we set Y to have a Poisson distribution, with $E(Y|A,V) := \exp(A + |V|)$. For data generating distribution 3, we take a distribution that is not in any generalized linear model family. More precisely, in data generating distribution 3, we set Y to be the sum of two independent random variables; the first has a Poisson distribution, with conditional mean given A, V set to $\exp(A + AV)$, and the second has two point masses with equal probabilities at 0 and 4 respectively. (We assume these two independent random variables are not observed, but that their sum, Y, is observed.) The log rate ratios are 1.5, 1, and $\approx$0.77 for data generating distributions 1, 2, and 3, respectively.

Table 1 below gives the mean squared error, relative efficiency, and coverage probabilities for our estimators, for each data generating distribution, and for sample sizes $n \in \{100, 500, 1000\}$. The mean squared error, observed relative

DOI: 10.2202/1557-4679.1138

efficiency, and coverage probabilities are calculated based on 10000 data sets generated according to each data generating distribution and sample size.

For all data generating distributions and all sample sizes considered, the targeted maximum likelihood estimator has smaller mean squared error than the unadjusted estimator. The relative efficiency in all cases is greater than 1, indicating greater precision for the targeted maximum likelihood estimator. The coverage probabilities are all close to the nominal 95%, being off by at most 3%. We note that for data generating distribution 1, if the working model is changed to include only main terms (and so becomes misspecified), the relative efficiencies decrease to approximately 1.25, 1.28, and 1.27,[6] for sample sizes $n = 100$, 500, and 1000, respectively; this is to be expected, since in general misspecification will lead to lower precision (but will still lead to consistent, asymptotically normal estimators).

We also ran a set of simulations using modifications of the data generating distributions 1, 2, and 3, where we decreased the magnitude of the parameter of interest (the log rate ratio). More precisely, we changed data generating distribution 1 to set Y to have a Poisson distribution, with $E(Y|A,V) := \exp(k(A + AV))$, for each of $k \in \{0.2, 0.4, 0.6, 0.8\}$. We changed data generating distribution 2 to set Y to have a Poisson distribution, with $E(Y|A,V) := \exp(kA+|V|)$, $k \in \{0.2, 0.4, 0.6, 0.8\}$. We changed data generating distribution 3 in an analogous way, setting Y to be the sum of two independent random variables; the first has a Poisson distribution, with conditional mean given A, V set to $\exp(k(A + AV))$, and the second has two point masses with equal probabilities at 0 and 4 respectively. Larger k correspond to larger values of the true log rate ratio. $k = 1$ would correspond to the original data generating distributions.

The results for these modified versions of data generating distributions 1, 2, and 3 are given in Table 2, where we focus on the relative efficiency, for clarity. For the modified version of data generating distributions 1 and 3, for all three sample sizes considered (n=100, 500, 1000), the relative efficiency grew as k got larger. For the modified version of data generating distribution 2, for each of the three sample sizes considered (n=100, 500, 1000), the relative efficiency remained roughly constant as k got larger (but differed by sample size); in this case, the asymptotic relative efficiency is 1, and so for larger sample sizes, the relative efficiency converges to 1.

[6]These values are not shown in Table 1, since all values in Table 1 are generated based on the targeted maximum likelihood estimator using the working model with main terms and an interaction term, as described earlier in this section.

Table 1: The Mean Squared Error (MSE), Relative Efficiency, and Coverage Probability of Nominal 95% Confidence Intervals (CI's) for the Targeted Maximum Likelihood Estimator (TMLE) vs. Unadjusted Estimator (UNADJ), at Sample Sizes n=100, 500, and 1000.

		MSE	Relative Efficiency	CI Coverage
Data Generating Distribution 1				
For n=100:	UNADJ	0.057	1.35	0.93
	TMLE	0.042		0.94
For n=500:	UNADJ	0.012	1.41	0.94
	TMLE	0.008		0.94
For n=1000:	UNADJ	0.006	1.42	0.94
	TMLE	0.004		0.94
Data Generating Distribution 2				
		MSE	Relative Efficiency	CI Coverage
For n=100:	UNADJ	0.045	1.10	0.93
	TMLE	0.041		0.92
For n=500:	UNADJ	0.009	1.02	0.95
	TMLE	0.009		0.95
For n=1000:	UNADJ	0.004	1.02	0.95
	TMLE	0.004		0.95
Data Generating Distribution 3				
		MSE	Relative Efficiency	CI Coverage
For n=100:	UNADJ	0.031	1.29	0.94
	TMLE	0.024		0.94
For n=500:	UNADJ	0.006	1.31	0.94
	TMLE	0.005		0.95
For n=1000:	UNADJ	0.003	1.31	0.95
	TMLE	0.002		0.95

DOI: 10.2202/1557-4679.1138

Table 2: The **Relative Efficiency** for the Targeted Maximum Likelihood Estimator (TMLE) vs. Unadjusted Estimator (UNADJ), at Sample Sizes n=100, 500, and 1000.

Data Generating Distribution 1					
	$k = 0.2$	$k = 0.4$	$k = 0.6$	$k = 0.8$	$k = 1$
For n=100:	1.00	1.04	1.11	1.23	1.35
For n=500:	1.01	1.04	1.12	1.26	1.41
For n=1000:	1.01	1.05	1.12	1.24	1.42
Data Generating Distribution 2					
	$k = 0.2$	$k = 0.4$	$k = 0.6$	$k = 0.8$	$k = 1$
For n=100:	1.09	1.09	1.09	1.10	1.10
For n=500:	1.03	1.02	1.02	1.03	1.02
For n=1000:	1.01	1.02	1.01	1.01	1.02
Data Generating Distribution 3					
	$k = 0.2$	$k = 0.4$	$k = 0.6$	$k = 0.8$	$k = 1$
For n=100:	1.00	1.01	1.05	1.14	1.29
For n=500:	1.00	1.02	1.06	1.15	1.31
For n=1000:	1.00	1.02	1.06	1.17	1.31

6 Brief Description of Targeted Maximum Likelihood Estimation

In the Appendix, we prove Theorem 1 using targeted maximum likelihood methodology. We give a brief overview here; a full description is given in (van der Laan and Rubin, 2006; van der Laan et al., 2009). Targeted maximum likelihood is a general methodology for estimation and inference. It can be used to estimate finite-dimensional, pathwise differentiable parameters, which include, for example, the following: marginal treatment effects (which are the parameters considered in this paper), the parameter of a marginal structural model, the parameter of a structural nested model, the parameter of a proportional hazards model, and the effect of static or dynamic treatments. Targeted maximum likelihood can also be used to estimate more general parameters including infinite-dimensional, non-pathwise differentiable parameters.

Targeted maximum likelihood estimation has several important advantages over standard maximum likelihood estimation and estimating function-based methodologies. When estimating parameters in the nonparametric model[7], maximum likelihood estimation based on assuming a parametric model (or based on selecting a parametric model using a sieve) may suffer severe bias due to model misspecification. A major improvement, especially in problems with high-dimensional confounding variables, is estimating function based methodology (Robins, 1986, 1987; van der Laan and Robins, 2002). However, these methods still have limitations. These include (1) in general not having a satisfactory way to deal with multiple solutions to an estimating equation, (2) only applying to problems that can be expressed in terms of a parameter of interest and a variation independent nuisance parameter, and (3) not being invariant to monotone transformations of the parameter of interest. Targeted maximum likelihood does not have any of these limitations. In addition, in many situations, targeted maximum likelihood can be simply implemented using standard statistical software.

6.1 Targeted Maximum Likelihood Algorithm

We now give a brief overview of the general algorithm for constructing the targeted maximum likelihood estimator. We follow this with several examples.

[7]By nonparametric model, we generally mean the model consisting of all continuous densities with respect to a given dominating measure. In this paper, we also use "nonparametric model" to describe the model that makes no assumptions on the density of the data generating distribution except that treatment A is randomized, so is independent of baseline variables V.

DOI: 10.2202/1557-4679.1138

The general idea in the algorithm is to start with an initial estimator for the density of the data generating distribution, and then make a series of updates that focus on improved estimation of the parameter of interest.[8] Heuristically, each update involves choosing a direction (score) for which the parameter of interest is most sensitive; the current estimate of the overall density is then pushed in this direction to the extent that the likelihood increases maximally. The final estimator for the parameter of interest is the plug-in (substitution) estimator of the density resulting from the last update. We give the basic skeleton of the targeted maximum likelihood algorithm. Improvements and extensions, such as collaborative targeted maximum likelihood (van der Laan and Gruber, 2009) and targeted loss-based learning (van der Laan and Rubin, 2006), build on this basic skeleton.

The targeted maximum likelihood algorithm takes as input the data, the assumed model $\mathcal{M}$, and the parameter ψ to be estimated (which formally is a function mapping densities in the model $\mathcal{M}$ to e.g. real-valued scalars or vectors). In the case considered in this paper, the observed data on a single subject consists of the vector (Y, A, V). The parameter of interest ψ is a smooth function, which we denote by r, of the mean outcomes given each of the two possible treatments: $E(Y|A = 0), E(Y|A = 1)$. As discussed above, different choices of the function r correspond to the parameter ψ being the risk difference, risk ratio, or log odds ratio, for example. The model $\mathcal{M}$ is nonparametric (making no assumptions), except for the assumption, ensured by randomization, that treatment assignment A and baseline variables V are independent.

For a given parameter of interest ψ, and model $\mathcal{M}$, the targeted maximum likelihood estimator is constructed in the following six steps:

1. An initial estimate p_0 of the density of the data generating distribution is constructed, by any method. For example, standard maximum likelihood estimation using a parametric working model (which we do not assume to be correctly specified) could be used to generate p_0. In the general targeted maximum likelihood algorithm, it is not necessary to estimate the full density, only the part on which the parameter ψ depends; for

[8]The motivation for the update step of the targeted maximum likelihood algorithm is related to the heuristic idea used in the one-step estimator for updating a given $\sqrt{n}$-consistent estimator in the direction of the efficient influence function (Bickel et al., 1993). However, a major difference is that in targeted maximum likelihood, the whole density is updated rather than just the parameter estimate, and the overall likelihood of the density estimate is increased at each iteration. While the one-step estimator may result in parameter estimates that do not correspond to any density in one's model, this will not happen for targeted maximum likelihood estimation.

simplicity, here we consider the case in which we estimate the full density. We use the term "density" below in the general sense[9], representing a continuous density for continuous data or a frequency function for discrete data; thus, the algorithm can be applied to situations involving both continuous variables and discrete variables.

2. The efficient influence function (also called the efficient influence curve) for the parameter ψ in the model $\mathcal{M}$ is computed, at p_0. Methods for finding the efficient influence function for a wide variety of parameters can be found in (van der Laan and Robins, 2002). The informal reasoning for focusing on the efficient influence function is that it gives the direction (score) for which the parameter is most sensitive to small fluctuations, to first order.

3. A parametric model with parameter vector ϵ and corresponding densities $\{p(\epsilon)\}$ is constructed that (i) equals the initial density p_0 at $\epsilon = 0$ and (ii) has score at $\epsilon = 0$ whose linear span contains the efficient influence function at p_0 (which was computed in step 2). The motivation is that we would like to construct a "least-favorable" model for our parameter, that is, a model that allows improvement in the direction in which the parameter we are estimating is most sensitive. Here ϵ is in general a real-valued vector, and we write $\epsilon = 0$ to mean that each component of this vector equals 0.

4. The parameter ϵ of the parametric model from the previous step is estimated using maximum likelihood estimation; call the resulting maximum likelihood estimator $\hat{\epsilon}$. This estimation step can be done, in many cases, by fitting a regression using standard software. We let p_1 denote the density $p(\hat{\epsilon})$. The density p_1 will have at least as large a likelihood as p_0, since by condition (i) of the previous step, p_0 is in this parametric model. An informal reason for desiring property (ii) in the previous step is that "optimal" estimators[10] solve an estimating equation involving the efficient influence function (Bickel et al., 1993; van der Vaart, 1998); iteration of steps 2-4 of this algorithm will lead to a final density approximately solving this estimating equation.

5. We then replace the initial density estimate p_0 by our new density p_1, and repeat steps 2-4 until the algorithm converges to a final density p (that is, until $||\hat{\epsilon}||$ is sufficiently small). In many cases, the algorithm

[9]The density can be with respect to any given measure (including counting measure).
[10]By "optimal" we mean semiparametric efficient estimators.

DOI: 10.2202/1557-4679.1138

will have converged (that is, $\hat{\epsilon} = 0$) after just a single iteration of steps 2-4.

6. Once the algorithm converges to a final density p, the targeted maximum likelihood estimator for the parameter ψ is the plug-in estimator of ψ at p. That is, we evaluate the parameter ψ at the final density p.

It is often useful to decompose the density for the observed data into components. For example, we can write the initial density $p_0(Y, A, V)$ as the product of $p_{01}(Y|A, V), p_{02}(A|V)$, and $p_{03}(V)$. One choice of initial estimator for the density of Y given A, V is the fit of a generalized linear model for Y given A, V. We give an example below.

6.2 Application of Targeted Maximum Likelihood Algorithm to Estimating Marginal Means in Randomized Trials

We now apply the above algorithm to a special case of our main result, in which Y is binary and a logistic regression working model is used. This example was given by Moore and van der Laan (2007), but we show it again here for completeness; in this paper we generalize some of the results in (Moore and van der Laan, 2007) to a variety of commonly used families of generalized linear models with canonical links. Also, for simplicity in this example, we let our parameter of interest be $E(Y|A = 1)$, which is the population mean were everyone assigned treatment $A = 1$. This corresponds to setting the function $r(x, y)$ in the definition of the parameter of interest to be $r(x, y) := y$. We will follow steps 1-5 of the targeted maximum likelihood algorithm to estimate the density $p(Y, A, V)$ and then, as in step 6, compute the plug-in estimator of our parameter $\psi := E(Y|A = 1)$. Though we use a binary outcome and a logistic regression working model here, the same steps lead to consistent, asymptotically normal, locally efficient estimators when using any of the generalized linear models with canonical links described in Section 2.3 as working models (e.g. Poisson regression with log link).

For step 1 of the algorithm, we separately specify initial estimators for the components of the density of (V, A, Y), denoted by $p_{01}(Y|A, V), p_{02}(A|V)$, and $p_{03}(V)$. We let our estimator $p_{01}(Y|A, V)$ be the maximum likelihood fit of the following logistic regression working model:

$$P(Y = 1|A, V) := \text{logit}^{-1}(\beta_0 + \beta_1 A + \beta_2 V + \beta_3 AV). \tag{10}$$

This is just one possible choice of terms for the model–any set of terms can be included, as long as an intercept and a main term for the treatment variable A

are included, as we discuss below. We fit the model with maximum likelihood estimation to produce $\hat{\beta}$, and set $p_{01}(Y|A,V)$ to be the density corresponding to the fit model:

$$p_{01}(Y=1|A,V) := \text{logit}^{-1}(\hat{\beta}_0 + \hat{\beta}_1 A + \hat{\beta}_2 V + \hat{\beta}_3 AV). \tag{11}$$

We have that $p_{02}(A|V)$ is known, and equals $1/2$, by randomization. Lastly, our estimator for $p_{03}(V)$ is the empirical distribution of V. Our overall initial density estimate is then the product $p_{01}(Y|A,V)p_{02}(A|V)p_{03}(V)$, which we denote by p_0.

Step 2 involves computing the efficient influence function for the parameter ψ. The efficient influence function for $P(Y=1|A=1)$, at a given density p, in our model only assuming independence of A, V is:

$$\frac{A(Y - p(Y=1|A=1,V))}{p(A=1)} + p(Y=1|A=1,V) - p(Y=1|A=1) \tag{12}$$

This efficient influence function is derived, for example, in (van der Laan and Robins, 2002). Note that by design we have $P(A=1) = 1/2$. Let $\hat{\psi}_0 := \frac{1}{n}\sum_{i=1}^{n} p_{01}(Y=1|A=1,V_i)$, which is the plug-in estimate at the initial density p_0; this will be used in step 3.

Step 3 involves choosing a parametric model $\{p(\epsilon)\}$ satisfying the conditions (i) and (ii) given in step 3 above. That is, we want to construct a model $\{p(\epsilon)\}$ that equals the initial density p_0 at $\epsilon = 0$, and (ii) has score at $\epsilon = 0$ whose linear span contains the efficient influence function (12) at p_0. We now give the general idea for how to build such a parametric model by adding a "clever covariate" (defined below) to the regression model (11). We will leverage a useful property of generalized linear models with canonical links.

Consider a generalized linear model for Y given A, V with canonical link g, as defined in (1), and with linear part η consisting of a single term $\epsilon C(A,V)$. The score (derivative of the log-likelihood) at $\epsilon = 0$ is equal to
$(Y - \mu(A,V))C(A,V)$, where $\mu(A,V)$ is the mean of Y given A, V, according to the generalized linear model. Setting $C(A,V) = A$, we have that the score at $\epsilon = 0$ is $A(Y - \mu(A,V))$, which has the same form as the first part of the efficient influence function (12). Thus, by setting our parametric model $\{p(\epsilon)\}$ to be the generalized linear model with linear part η consisting of a single term $\epsilon C(A,V)$, the score at $\epsilon = 0$ will be the first part of the efficient influence function (12) at p_0. An extension of this procedure, described in detail below, results in a model with score at $\epsilon = 0$ equal to the entire efficient influence function (12) at p_0. Such a model satisfies condition (ii) given in step 3 above; we can make it satisfy condition (i) by adding an offset term to the model (described below).

DOI: 10.2202/1557-4679.1138

We call a choice $C(A,V)$ of a term in a generalized linear model a "clever covariate" if it results in a score that equals a part of the efficient influence function for the given problem. Methods for obtaining clever covariates for a variety of parameters and models are given in (van der Laan and Rubin, 2006; Moore and van der Laan, 2007; Polley and van der Laan, 2009; van der Laan et al., 2009). After a clever covariate for a given problem is derived, estimation then involves simply refitting a regression including that covariate and with the linear part of the initial regression as an offset, as we describe next.

We now build a parametric model $\{p(\epsilon)\}$ satisfying the conditions (i) and (ii) given in step 3 above. Define "clever covariates" $C_1(A) := A$, and $C_2(V) := p_{01}(Y=1|A=1,V) - \hat{\psi}_0$. For each $\epsilon = (\epsilon_1, \epsilon_2)$, define the density in the parametric model $p(\epsilon)(Y,A,V) := p_1(\epsilon)(Y|A,V)p_2(\epsilon)(A|V)p_3(\epsilon)(V)$, where

$$p_1(\epsilon)(Y=1|A,V) := \text{logit}^{-1}(\hat{\beta}_0 + \hat{\beta}_1 A + \hat{\beta}_2 V + \hat{\beta}_3 AV + \epsilon_1 C_1(A)), \quad (13)$$

$$p_2(\epsilon)(A=1|V) := 1/2, \quad (14)$$

$$p_3(\epsilon)(V) := s_{\epsilon_2} \exp(\epsilon_2 C_2(V)) p_{03}(V), \quad (15)$$

where the constant $s_{\epsilon_2} := 1/[\frac{1}{n}\sum_{i=1}^n \exp(\epsilon_2 C_2(V_i))]$ is chosen so that $p_3(\epsilon)(v)$ integrates to 1. In this definition of the model $\{p(\epsilon)\}$, we consider $\hat{\beta}$ and $\hat{\psi}_0$ as fixed numbers (having been computed in step 1). It follows that conditions (i) and (ii) hold for this model, since substituting 0 for each component of ϵ results in the initial density p_0, and the components of the score at $\epsilon = (0,0)$ (which we sometimes write more concisely as $\epsilon = 0$) are:

$$\begin{aligned} \frac{d}{d\epsilon_1}[\log p(\epsilon)(Y,A,V)]|_{\epsilon=0} &= \frac{d}{d\epsilon_1}[\log p_1(\epsilon)(Y|A,V)]|_{\epsilon=0} \\ &= (Y - p_{01}(Y=1|A=1,V))C_1(A), \end{aligned} \quad (16)$$

and

$$\frac{d}{d\epsilon_2}[\log p(\epsilon)(Y,A,V)]|_{\epsilon=0} = \frac{d}{d\epsilon_2}[\log p_3(\epsilon)(V)]|_{\epsilon=0} = C_2(V). \quad (17)$$

Thus, substituting the definitions of $C_1(A)$ and $C_2(V)$, we have that the linear span of these components of the score at $\epsilon = 0$ includes the efficient influence function at p_0 as given in (12).

Step 4 involves computing the maximum likelihood estimator $(\hat{\epsilon}_1, \hat{\epsilon}_2)$ for the parameter (ϵ_1, ϵ_2) of the model defined in (13), (14), (15). To get $\hat{\epsilon}_1$, we fit the logistic regression (13), where we enter the expression $\hat{\beta}_0 + \hat{\beta}_1 A + \hat{\beta}_2 V + \hat{\beta}_3 AV$ in (13) as an offset (since we consider $\hat{\beta}$ to be fixed), so that only ϵ_1 can be varied. Since we already had the term A in the original logistic regression for p_{01}, and since $C_1(A) = A$, we must have $\hat{\epsilon}_1 = 0$. To get $\hat{\epsilon}_2$, we first note that

since by our having chosen $p_{03}(V)$ to be the empirical distribution of V and by the definition of $\hat{\psi}_0$ from step 2, we have

$$\begin{aligned}\sum_{i=1}^{n} \frac{d}{d\epsilon_2}[\log p(\epsilon)(Y_i, A_i, V_i)]|_{\epsilon=(0,0)} &= \sum_{i=1}^{n} C_2(V_i) \\ &= \sum_{i=1}^{n}[p_{01}(Y=1|A=1,V_i) - \hat{\psi}_0] \\ &= 0.\end{aligned}$$

Also, for all v, $p_3(\epsilon)(v)$ is a strictly concave function of ϵ_2, which follows directly from its definition. Thus, the maximizer of the log-likelihood occurs at $\hat{\epsilon}_2 = 0$. Putting this all together, we have $(\hat{\epsilon}_1, \hat{\epsilon}_2) = (0, 0)$, and our updated density $p(\hat{\epsilon}) = p(\mathbf{0})$, which is precisely the initial density from step 1.

Step 5 involves iteration of the steps 2-4, until $\hat{\epsilon}$ is sufficiently small. Since it's already equal to $(0, 0)$ after the first iteration, we need not do any more iterations.

Lastly, for step 6, we compute the plug-in estimate of ψ, that is, ψ evaluated at the final density p output from step 5. Thus, our estimate of ψ is

$$\begin{aligned}\hat{\psi} &= \frac{1}{n}\sum_{i=1}^{n} p_{01}(Y=1|A=1,V_i) \\ &= \frac{1}{n}\sum_{i=1}^{n} \text{logit}^{-1}(\hat{\beta}_0 + \hat{\beta}_1 + \hat{\beta}_2 V_i + \hat{\beta}_3 V_i). \qquad (18)\end{aligned}$$

We have shown that for the parameter, model, and observed data given at the beginning of this section, the targeted maximum likelihood algorithm reduces to fitting the logistic regression (11) and then computing the empirical mean of the resulting fit, setting $A = 1$, as in (18). Certain targeted maximum likelihood estimators, for many parameters and many models, can be shown to converge in just one iteration, and will involve the use of clever covariates in regression fits. This allows the construction of simple to compute estimators that can be shown to have desirable properties (e.g. the double robustness property described in (van der Laan and Robins, 2002)).

It will not always be the case, as above, that the clever covariate will already have been included as a term in the regression; in this case, one simply fits the regression with added term (which we called the "clever covariate"), fixing the previously computed coefficients (by using an offset when fitting the regression). We give an example of this situation next.

DOI: 10.2202/1557-4679.1138

6.3 Example of Targeted Maximum Likelihood Convergence Requiring More than One Iteration

In general, when using generalized linear working models with *non-canonical* links in the targeted maximum likelihood algorithm, convergence will not occur in a single iteration. For example, using a probit regression working model (which uses the inverse normal cumulative distribution function as link) above would require multiple iterations. This occurs because the "clever covariate" added to the regression will depend on parts of the density estimate that are updated at each iteration. This is in contrast to the case of generalized linear models with canonical links (as above), where convergence occurs in a single iteration of the targeted maximum likelihood algorithm. We give more details of a targeted maximum likelihood estimator using probit regression in Section 8.3 of the Appendix.

7 Discussion

We described a large class of easy to implement, model-based estimators that leverage baseline variables to improve precision of estimates of marginal effects in randomized trials. These estimators are guaranteed to be asymptotically unbiased and asymptotically normal, even when the working models used are arbitrarily misspecified. When the working model is correctly specified, the estimators are efficient. Our results demonstrate one application of targeted maximum likelihood methodology.

Though the focus of this paper was on randomized trials, the targeted maximum likelihood estimator can also be used to estimate treatment effects in observational studies. For example, consider estimating the marginal mean $E(Y|A=1)$ using a logistic regression working model, as in Section 6.2. There, the "clever covariate" used was $C_1(A,V) := A$. In an observational study, we would use as "clever covariate" $C_1(A,V) := A/\hat{p}(A|V)$ where $\hat{p}(A|V)$ is an estimate of the probability of A given V. Note that this differs from standard inverse probability weighting methods, which instead of incorporating weights within terms in the regression model, weight the overall regression. For more examples of targeted maximum likelihood estimators for a wide variety of parameters and models, see e.g. (van der Laan and Rubin, 2006; Moore and van der Laan, 2007; Polley and van der Laan, 2009; van der Laan et al., 2009).

8 Appendix

In this appendix, we first prove Theorem 1. We next give R code for the estimator from Section 3. We give more details and R code for the example of a targeted maximum likelihood estimator requiring more than one iteration of the targeted maximum likelihood algorithm from Section 6.3. Lastly, in Section 8.4, we briefly discuss incorporating model selection into the estimation method we presented in this paper

8.1 Proof of Theorem 1

We prove Theorem 1. First, we show that the targeted maximum likelihood estimator in our setting is of the simple form given in the beginning of Section 3. Next, we verify that the regularity conditions given in Section 2 are sufficient to prove all the claims in Theorem 1.

Consider the model used throughout this paper, where the data consist of i.i.d. observations (V_i, A_i, Y_i) and the randomized treatment A_i is assumed to take values 0 and 1 with probability 1/2, independent of the baseline variables V_i. The parameter being estimated is a smooth function r of the conditional means $E(Y|A = 0)$ and $E(Y|A = 1)$. The efficient influence function for this parameter in the nonparametric model is then a linear combination of the efficient influence functions for the conditional means $E(Y|A = 0)$ and $E(Y|A = 1)$. At any given density p, these efficient influence functions are given by (4), (5) above.

We will use a generalized linear model with canonical link, as described in Section 2. We use the notation from Section 2, and make all the assumptions from Section 2. We note that under the assumptions in Section 2 on our families of generalized linear models with canonical links, we have $\ddot{b}(\eta) := \frac{d^2b}{d\eta^2} > 0$ for all η.

We first extract some useful information from the fact that $\hat{\beta}$ is the maximum likelihood estimator of the generalized linear model defined above. Let $p_{01}(Y|A, V)$ denote the the maximum likelihood estimate for the density of Y given A, V, using the above generalized linear model. Under the regularity assumptions made in Section 2, we have that the derivative of the log-likelihood at $\hat{\beta}$ must be 0. The derivative of the log of (1) is $(\partial\eta/\partial\beta)(Y - \dot{b}(\eta)) = (\partial\eta/\partial\beta)(Y - E_{p_{01}}(Y|A_i, V_i)))$, based on the fact for generalized linear models that $\mu(A, V) = \dot{b}(\eta(A, V))$. Since we assumed the linear part η of the generalized linear model contains an intercept term and also contains A as a main

DOI: 10.2202/1557-4679.1138

term, this implies

$$\sum_{i=1}^{n}(Y_i - E_{p_{01}}(Y|A_i, V_i)) = 0, \tag{19}$$

and

$$\sum_{i=1}^{n} A_i(Y_i - E_{p_{01}}(Y|A_i, V_i)) = 0. \tag{20}$$

The targeted maximum likelihood algorithm requires an initial density estimator p_0 for the data generating distribution of (V, A, Y). It will be based on the maximum likelihood estimate $\hat{\beta}$ from the generalized linear model and the set of observed baseline variables $\{V_i\}$. We set

$$p_0(V, A, Y) = p_{01}(Y|A, V)p_{02}(A|V)p_{03}(V), \tag{21}$$

where we have

- $p_{01}(Y|A, V)$ is the maximum likelihood estimate for the density of Y given A, V, using the pre-specified generalized linear model,
- $p_{02}(A|V) = 1/2$, to reflect the known randomization probabilities, and
- $p_{03}(V)$ is the empirical distribution of V.

Since $p_{03}(V)$ was chosen to be the empirical distribution of V, and by our choice of $p_{02}(A|V) = 1/2$, we have

$$\begin{aligned}
&\sum_{i=1}^{n}(E_{p_0}(Y|A = 1, V_i) - E_{p_0}(Y|A = 1)) \\
&= \sum_{i=1}^{n} E_{p_{01}}(Y|A = 1, V_i) - \sum_{i=1}^{n}[(1/n)\sum_{j=1}^{n} E_{p_{01}}(Y|A = 1, V_j))] \\
&= 0.
\end{aligned} \tag{22}$$

Similarly, we have

$$\begin{aligned}
&\sum_{i=1}^{n}(E_{p_0}(Y|A = 0, V_i) - E_{p_0}(Y|A = 0)) \\
&= \sum_{i=1}^{n} E_{p_{01}}(Y|A = 0, V_i) - \sum_{i=1}^{n}[(1/n)\sum_{j=1}^{n} E_{p_{01}}(Y|A = 0, V_j))] \\
&= 0.
\end{aligned} \tag{23}$$

We now define our parametric model $\{p(\epsilon)\}$ that satisfies conditions (i) and (ii) of step 3 of the targeted maximum likelihood algorithm outlined in Section 6. It will involve adding a term to the linear part of the generalized linear model and also modifying $p_{03}(V)$. We let $p(\epsilon)$ be defined as $p_{01,\epsilon}(Y|A,V)p_{02}(A|V)p_{03,\epsilon}(V)$, where $\epsilon = (\epsilon_1, \epsilon_2, \epsilon_3, \epsilon_4)$, for densities $p_{01,\epsilon}(Y|A,V)$ and $p_{03,\epsilon}(V)$ defined next. First, $p_{01,\epsilon}(Y|A,V)$ is defined in terms of the generalized linear model as $\exp(Y\eta' - b(\eta') + c(Y,\phi))$, where $\eta' = \hat{\eta} + \epsilon_1 + \epsilon_2 A$, and $\hat{\eta} = \sum_j \hat{\beta}_j f_j(A,V)$. We note that $\dot{b}(\hat{\eta}(A,V)) = E_{p_{01}}(Y|A,V)$, which follows from the fact that for any generalized linear model, $\dot{b}(\eta(A,V))$ is the mean of Y given A,V according to the model at β, which is proved in (Bickel and Doksum, 2001).

Next, we define

$$\begin{aligned} p_{03,\epsilon}(V) &= C_\epsilon \exp(\epsilon_3(E_{p_0}(Y|A=0,V) - E_{p_0}(Y|A=0)) \\ &\quad +\epsilon_4(E_{p_0}(Y|A=1,V) - E_{p_0}(Y|A=1)))p_{03}(V), \end{aligned} \tag{24}$$

where C_ϵ is chosen so that $p_{03,\epsilon}(v)$ integrates to 1.

Then $p(\epsilon)$ at $\epsilon = 0$ equals the initial density estimator p_0, and the components of the score of $p(\epsilon)$ at $\epsilon = 0$ equal

$$\frac{d}{d\epsilon_1}[\log p(\epsilon)]|_{\epsilon=0} = \frac{d}{d\epsilon_1}[\log p_{01,\epsilon}(Y|A,V)]|_{\epsilon=0} = (Y - \dot{b}(\hat{\eta})) = (Y - E_{p_{01}}(Y|A,V)), \tag{25}$$

$$\begin{aligned} \frac{d}{d\epsilon_2}[\log p(\epsilon)]|_{\epsilon=0} = \frac{d}{d\epsilon_2}[\log p_{01,\epsilon}(Y|A,V)]|_{\epsilon=0} &= A(Y - \dot{b}(\hat{\eta})) \\ &= A(Y - E_{p_{01}}(Y|A,V)) \end{aligned} \tag{26}$$

$$\frac{d}{d\epsilon_3}[\log p(\epsilon)]|_{\epsilon=0} = \frac{d}{d\epsilon_3}[\log p_{03,\epsilon}(V)]|_{\epsilon=0} = E_{p_0}(Y|A=0,V) - E_{p_0}(Y|A=0), \tag{27}$$

$$\frac{d}{d\epsilon_4}[\log p(\epsilon)]|_{\epsilon=0} = \frac{d}{d\epsilon_4}[\log p_{03,\epsilon}(V)]|_{\epsilon=0} = E_{p_0}(Y|A=1,V) - E_{p_0}(Y|A=1). \tag{28}$$

Thus, the efficient influence functions for $E(Y|A=0)$ and for $E(Y|A=1)$, (4) and (5) above, are in the linear span of the score of $p(\epsilon)$ at $\epsilon = 0$; this satisfies requirement (ii) in step 3 of the targeted maximum likelihood procedure given in Section 6.

We now show that the maximum likelihood estimator of ϵ for the model $\{p(\epsilon)\}$, is 0, whenever the conditions of Theorem 1 hold. By our assumption that the expected log-likelihood has a unique maximizer and the other assumptions in Section 2, we have that for sufficiently large n, the log-likelihood

DOI: 10.2202/1557-4679.1138

$\sum_{i=1}^{n} \log p(\epsilon)(V_i, A_i, Y_i)$ has a unique maximizer. By strict concavity of the log-likelihood (as proved in (Rosenblum and van der Laan, 2009, Appendix D) for our families of generalized linear models with canonical links), the maximum likelihood estimator $\hat{\epsilon}$ is the unique value of ϵ for which $d/d\epsilon[\sum_{i=1}^{n} \log p(\epsilon)(V_i, A_i, Y_i)] = 0$. Equations (19-23) and (25-28) imply $d/d\epsilon[\sum_{i=1}^{n} \log p(\epsilon)(V_i, A_i, Y_i)] = 0$ at $\epsilon = 0$, and so $\hat{\epsilon} = 0$ is the maximum likelihood estimator for the model $\{p(\epsilon)\}$. Therefore, the targeted maximum likelihood procedure converges in zero steps. Furthermore, since the final density output by the targeted maximum likelihood algorithm is equal to the initial density estimator p_0, we have that the targeted maximum likelihood estimator of the parameter $(E(Y|A = 0), E(Y|A = 1))$ is exactly as given in Theorem 1.

Theorem 1 requires the existence of a maximizer β^* of the expected log-likelihood $E(Y\eta - b(\eta) + c(Y, \phi))$, where the expectation is with respect to the data generating distribution. Given the assumptions in Section 2, this is sufficient to ensure that the the maximum likelihood estimator $\hat{\beta}_n$ converges to β^* and that $\sqrt{n}(\hat{\beta}_n - \beta^*)$ is asymptotically normal. This follows from the strict concavity of the expected log-likelihood for generalized linear models with canonical links, proved in (Rosenblum and van der Laan, 2009, Appendix D).

So far we have shown that the targeted maximum likelihood estimator in our setting is of the simple form given in Section 3. We now verify that the regularity conditions given in Section 2 are sufficient to prove all the claims in Theorem 1. To this end, we apply Theorem 1 of van der Laan and Rubin (2006), which under conditions that we verify below, gives that the estimator $r(\hat{E}_0, \hat{E}_1)$ is asymptotically unbiased with asymptotic variance as defined in (6) in Section 4, and is locally efficient.

There are five conditions in Theorem 1 of van der Laan and Rubin (2006), which we verify now. We note that we apply Theorem 1 of van der Laan and Rubin (2006) to the parameter (E_0, E_1) and estimator $(\hat{E}_0, \hat{E}_1)$. This then implies the desired results for the parameter $r(E_0, E_1)$ and estimator $r(\hat{E}_0, \hat{E}_1)$. Denote the density p_0 defined above, at sample size n, by p_0^n. To apply Theorem 1 of van der Laan and Rubin (2006), we need to show:

i. The model is convex.

ii. The parameter (E_0, E_1) is linear.

iii. Our estimator $(\hat{E}_0, \hat{E}_1)$ of (E_0, E_1) satisfies

$$(\hat{E}_0, \hat{E}_1) - (E_0, E_1) = \frac{1}{n} \sum_{i=1}^{n} D(p_0^n)(V_i, A_i, Y_i) - E_{p^*} D(p_0^n)(V, A, Y),$$

where E_{p^*} is the expectation over the variables V, A, Y with respect to the data generating distribution, and where p_0^n is considered fixed.

iv. $D(p_0^n)$ is in a Donsker class with probability tending to 1.

v. $E_{p^*}\left(D_i(p_0^n)(V,A,Y) - D_i(p(\beta^*))(V,A,Y)\right)^2$ converges to 0 in probability, for $i \in \{1,2\}$, where D_1, D_2 are defined in (4) and (5).

Proof of conditions (i)-(v) above:

Condition (i) follows from our model being nonparametric except for assuming, due to randomization, that $p(A|V) = 1/2$.

Condition (ii) follows since for p^1, p^2 two densities in our model, and defining $p^3 = \lambda p^1 + (1-\lambda)p^2$, for $\lambda \in [0,1]$, we have

$$\begin{aligned} p^3(Y|A) &= p^3(Y,A)/p^3(A) \\ &= \int p^3(v,A,Y)dv/(1/2) \\ &= \lambda \int p^1(v,A,Y)dv/(1/2) + (1-\lambda)\int p^2(v,A,Y)dv/(1/2) \\ &= \lambda p^1(Y|A) + (1-\lambda)p^2(Y|A). \end{aligned}$$

Thus, the conditional mean of Y given A under p^3 is the convex combination of these conditional means under p^1 and p^2, which proves linearity of the parameter (E_0, E_1) in our model.

Condition (iii) follows since $\frac{1}{n}\sum_{i=1}^n D(p_0^n)(V_i, A_i, Y_i) = 0$ using the definitions (4), (5) and applying (19), (20), (22), and (23), and since

$$\begin{aligned} E_{p^*}D_1(p_0^n)(V,A,Y) &= E_{p^*}(1-A)(Y - E_{p_0^n}(Y|A=0,V))/p_0^n(A=0) \\ &\quad + E_{p_0^n}(Y|A=0,V) - E_{p_0^n}(Y|A=0) \\ &= E_{p^*}(1-A)(Y)/(1/2) - E_{p^*}E_{p_0^n}(Y|A=0,V) \\ &\quad + E_{p^*}E_{p_0^n}(Y|A=0,V) - E_{p^*}E_{p_0^n}(Y|A=0) \\ &= E_{p^*}(1-A)(Y)/(1/2) - E_{p^*}E_{p_0^n}(Y|A=0) \\ &= E_{p^*}(Y|A=0) - \hat{E}_0 \\ &= E_0 - \hat{E}_0 \end{aligned}$$

where the second equality follows using the fact that A and V are independent in all of our densities p_0^n, and the second to last equality follows from E_{p^*} being with respect to V, A, Y and treating p_0^n as fixed; a similar derivation shows the analogous statement for $E_{p^*}D_2(p_0^n)(V,A,Y)$.

DOI: 10.2202/1557-4679.1138

To show (iv), first let $\mu_\beta(a,v)$ denote the mean of Y given $A=a, V=v$ according to the generalized linear model (1), which depends on β through the linear part $\eta = \sum_j \beta_j f_j(A,V)$. We will show the class of functions $\{\bar{D}_{\alpha_1,\alpha_2,\beta}(v,a,y)\}$ is Donsker, where we define

$$\begin{aligned}\bar{D}_{\alpha_1,\alpha_2,\beta}(v,a,y) &= ((1-a)(y-\mu_\beta(0,v))/(1/2)+\mu_\beta(0,v)-\alpha_1, \\ &\qquad a(y-\mu_\beta(1,v))/(1/2)+\mu_\beta(1,v)-\alpha_2), \end{aligned} \quad (29)$$

and we require $|\alpha_1| \leq M, |\alpha_2| \leq M, \beta \in B$, for B the set of possible β, defined in Section 2,[11] which was selected to ensure our class is over a bounded parameter set; additionally, our assumptions on boundedness of variables and the functions f_j in Section 2, combined with the functional forms of the generalized linear models we are considering, guarantee that the first and second derivatives of $\bar{D}$ with respect to v,a,y are uniformly bounded. We can then apply the result from (van der Vaart, 1998, Example 19.9, page 272), which implies this class of functions is a Donsker class. Since $D(p_0^n)$ are all contained in this class, condition (iv) above is satisfied. We note that Theorem 1 of van der Laan and Rubin (2006) only requires conditions (i) to (iv) in order to prove consistency of the estimator $(\hat{E}_0, \hat{E}_1)$, which we'll use below in proving (v).

Lastly, we show (v) above holds. Let $p(\beta^*)$ denote the density of Y given A,V defined in (1) corresponding to $\beta = \beta^*$, for β^* the maximizer of the expected log-likelihood $E(Y\eta - b(\eta) + c(Y,\phi))$, where the expectation is with respect to the data generating distribution p^*. Note that whenever the model (1) is misspecified, p^* and $p(\beta^*)$ will be different densities (where the former is the true data generating distribution and the latter is the projection under Kullback-Leibler divergence of the true data generating distribution on the working model); our theorem still holds in this case. Below we let $\hat{\beta}_n$ denote the maximum likelihood estimator from the generalized linear model fit at sample size n. We then have the following chain of inequalities, where we let

[11] In Section 2, we assumed all components of β must have absolute value at most M for some constant M, and for the Gamma and Inverse Normal families, B is further restricted to contain only β for which all components are positive and more than δ for some $\delta > 0$.

$p_{02}(A|V) = 1/2$ and $p^*_{03}(V)$ be the true density of V:

$$\begin{aligned}
& E_{p^*}\left(D_1(p_0^n)(V,A,Y) - D_1(p(\beta^*)p_{02}p^*_{03})(V,A,Y)\right)^2 \\
= & \int \{(1-a)(y-\mu_{\hat{\beta}_n}(0,v))/(1/2) + \mu_{\hat{\beta}_n}(0,v) - \hat{E}_0 \\
& \quad -[(1-a)(y-\mu_{\beta^*}(0,v))/(1/2) + \mu_{\beta^*}(0,v) - E_0]\}^2 p^*(v,a,y)dvdady \\
= & \int \left((1-2a)(\mu_{\beta^*}(0,v) - \mu_{\hat{\beta}_n}(0,v)) + E_0 - \hat{E}_0\right)^2 p^*(v,a)dvda \\
\leq & \; 2\int \left([\mu_{\beta^*}(0,v) - \mu_{\hat{\beta}_n}(0,v)]^2 + [E_0 - \hat{E}_0]^2\right) p^*(v,a)dvda \qquad (30) \\
= & \; 2\int \left([\mu_{\beta^*}(0,v) - \mu_{\hat{\beta}_n}(0,v)]^2 + [E_0 - \hat{E}_0]^2\right) p^*(v)dv \qquad (31) \\
\leq & \; C_1||\beta^* - \hat{\beta}_n||^2 + 2[E_0 - \hat{E}_0]^2. \qquad (32) \\
& \qquad (33)
\end{aligned}$$

where the the first equality follows from definitions; the second equality follows from canceling terms and noting that there is no longer any dependence on y; the inequality (30) follows from the bound $(x+y)^2 \leq 2(x^2+y^2)$ and noting that $1-2a$ is always either 1 or -1; the equality (31) follows from noting that there is no longer dependence on a; and the last line follows from $\mu_\beta(0,v)$ having first derivative uniformly bounded by a constant. The last line of the above display converges to 0 in probability since by our assumptions in Section 2, $\hat{\beta}_n$ converges to β^* in probability and also the consistency of $\hat{E}_0$ follows from conditions (i)-(iv) above, as described in Theorem 1 of van der Laan and Rubin (2006). An analogous bound as just derived proves that $E_{p^*}\left(D_2(p_0^n))(V,A,Y) - D_2(p(\beta^*))(V,A,Y)\right)^2$ converges to 0 in probability.

This completes our verification of the conditions (i)-(v) above of Theorem 1 of van der Laan and Rubin (2006), which implies that the estimator $(\hat{E}_0, \hat{E}_1)$ converges to $(E(Y|A=0), E(Y|A=1))$, and that $\sqrt{n}[(\hat{E}_0, \hat{E}_1) - (E(Y|A=0), E(Y|A=1))]$ is asymptotically normal with variance given by (6), and is locally efficient in that if the generalized linear model is correctly specified, then (6) achieves the efficiency bound for the nonparametric model. This completes the proof of Theorem 1.

□

Since the true data generating distribution is unknown, one a priori does not know whether there exists a maximizer β^* of the expected log-likelihood $E(Y\eta - b(\eta) + c(Y,\phi))$, where the expectation is taken with respect to the data generating distribution. However, as proved in (Rosenblum and van der Laan,

DOI: 10.2202/1557-4679.1138

2009, Appendix D), by strict concavity of the $E(Y\eta - b(\eta) + c(Y,\phi))$, we always have either (1) there is a unique maximizer of the expected log-likelihood or (2) the Euclidean norm of the maximum likelihood estimator grows without bound as sample size goes to infinity. Thus, as sample size n tends to infinity, with probability tending to 1, one will detect case (2), based on whether $\hat{\beta}_n$ exceeds the pre-specified threshold M from Section 2.

8.2 R Code to Compute Estimator from Section 3

We now give R code that computes the estimator given just before Theorem 1. The code below corresponds to the specific example of a Poisson model with log link and linear part $\beta_0 + \beta_1 A + \beta_2 V + \beta_3 AV$.

```
# Given vectors V, A, Y of length n containing baseline variables,
# treatment assignment and outcome, respectively,
# compute the estimated log rate ratio
modelfit <- glm(Y ~ 1 + A + V+ A*V,family=poisson)
E_0_hat <- mean(predict.glm(modelfit, type = "response",
  newdata=data.frame(A=rep(0,n),V=V)))
E_1_hat <- mean(predict.glm(modelfit, type = "response",
   newdata=data.frame(A=rep(1,n),V=V)))
log_rate_ratio_estimate <- log(E_1_hat/E_0_hat)
```

8.3 Details for Example of Targeted Maximum Likelihood Convergence Requiring More than One Iteration

As noted in Section 6.3, in general, when using generalized linear models with non-canonical links in the targeted maximum likelihood algorithm, convergence will not occur in a single iteration. For example, if the parameter of interest $\psi = E(Y|A=1)$ is the same as defined in Section 6.2, and a working model based on probit regression is used instead of logistic regression, this would require multiple iterations. This occurs because the "clever covariate" added to the regression will depend on parts of the density estimate that are updated at each iteration. We now give an example of a "clever covariate" that satisfies conditions (i) and (ii) of step 3 of the targeted maximum likelihood algorithm given in Section 6.1, for estimating the parameter defined in Section 6.2, but using probit regression as the working model.

We follow steps 1 and 2 of the procedure given in Section 6.2, except replacing the logistic regression working model by a probit regression working model. For step 3, we will have to choose a different "clever covariate" than $C_1(A)$ (but we still use clever covariate $C_2(V)$). This is because the score at $\epsilon = 0$ as computed at (16), if we had used the probit instead of the logistic link, would be the product

$$\begin{aligned} & C_1(A)(Y - p_{01}(Y = 1|A = 1, V)) \\ & \times(1/p_{01}(Y = 1|A = 1, V) + 1/p_{01}(Y = 0|A = 1, V)) \\ & \times\phi(\Phi^{-1}(p_{01}(Y = 1|A = 1, V))). \end{aligned}$$

where ϕ and Φ are the density and cumulative distribution function, respectively, for the standard normal distribution. One option then, for a "clever covariate" satisfying (i) and (ii) of step 3 of the algorithm from Section 6.1, is the following: let

$$\begin{aligned} C_1'(A, V) \;:=\; & A/\left[(1/p_{01}(Y = 1|A = 1, V) + 1/p_{01}(Y = 0|A = 1, V))\right. \\ & \left.\phi(\Phi^{-1}(p_{01}(Y = 1|A = 1, V)))\right]. \end{aligned} \tag{34}$$

Below is R code for the probit regression case, using the "clever covariate" (34) just described:

```
# Given vectors V, A, Y of length n containing baseline variables,
# treatment assignment
# and outcome, respectively, compute the estimator
# just described for E(Y|A=1)
# Define "clever covariate" C_1 as function of A and
# predicted probabilities of Y=1 given A=1, V for each subject
C_1_constructor <- function(A_values, Y_probs)
{return(ifelse((Y_probs==0 | Y_probs==1),0,A_values/
((1/Y_probs + 1/(1-Y_probs))*(dnorm(qnorm(Y_probs))))))}
# Step 1: Initial model fit
initial_modelfit <- glm(Y ~ 1 + A + V+ A*V,
   family=binomial(link=probit))
# Initialize vector to store value of "Linear Part" of Model
# to be used as the offset for the model update
# (step 4 of the algorithm from Section 3)
# This is equivalent to eta(A,V) as defined in Section 2
linear_part <- predict.glm(initial_modelfit,
   newdata=data.frame(A,V))
# Define analogous vector to store value of "Linear Part"
```


DOI: 10.2202/1557-4679.1138

```
# Setting A=1. This is equivalent to eta(1,V)
linear_part_A_set_to_1 <- predict.glm(initial_modelfit,
     newdata=data.frame(A=1,V=V))
# Next, iterate steps 2-4 of the targeted maximum likelihood
# algorithm from Section 3 until
# epsilon_hat is close to 0:
epsilon_hat <- 1
while(abs(epsilon_hat) > 0.0001)
{
# Construct "Clever Covariate" C_1
C_1 <- C_1_constructor(A,
pnorm(linear_part_A_set_to_1))
# Update model fit using "Clever Covariate" C_1
modelfit <- glm(Y ~ offset(linear_part) -1 +C_1,
family=binomial(probit))
epsilon_hat <<- modelfit$coefficients[1]
# Update vectors storing linear parts for each subject
linear_part <- linear_part + epsilon_hat*C_1
linear_part_A_set_to_1 <- linear_part_A_set_to_1 +
epsilon_hat*C_1_constructor(rep(1,length(A)),
pnorm(linear_part_A_set_to_1))
}
E_1_hat <- mean(pnorm(linear_part_A_set_to_1))
```

8.4 Incorporating Model Selection

One of the assumptions in Theorem 1 (and listed in Section 2.3 of the paper) is that the working model used is fixed; that is, the working model does not change with sample size and is selected prior to looking at the data. In practice this is a restrictive assumption, and we briefly describe how this assumption can be relaxed to allow some forms of model selection.

Consider a list of K working models, each satisfying the conditions of Theorem 1. Here, we restrict attention to the case where this list of working models is pre-specified (before looking at the data), and does not change with sample size n. As an example of such a list of working models, one could pre-specify three log-linear working models: the first with only the intercept and treatment variable as terms, the second having these and the baseline variable V as main terms, and the third having these terms and all interaction terms. We refer to these below as working models 1, 2, and 3, respectively. We require that an algorithm for selecting among the list of K working models be pre-

specified (though the algorithm may take the data as input). The estimator for the parameter $r(E(Y|A=0), E(Y|A=1))$ is now computed in two steps: first, we use a model selection algorithm to choose a working model; second, we use this working model in the construction given before Theorem 1 to compute our estimate of $r(E(Y|A=0), E(Y|A=1))$.

We now give an example of an algorithm for selecting among working models 1, 2, and 3 that results in consistent, asymptotically normal estimators for the parameter $r(E(Y|A=0), E(Y|A=1))$. The algorithm is based on the idea of empirical efficiency maximization (Rubin and van der Laan, 2008); this involves computing, for each working model in the list, the corresponding estimated asymptotic variance of $r(\hat{E}_0, \hat{E}_1)$, as given in (7) below. We then select the working model for which this estimated asymptotic variance is smallest. Here, we assume that there is a unique model in the list of K working models having the smallest corresponding asymptotic variance.[12] We show in the Appendix that for any working model satisfying the conditions of Theorem 1, the estimated asymptotic variance (7) converges to the true asymptotic variance, as sample size tends to infinity. It then follows that for large enough sample size, the working model with corresponding smallest asymptotic variance will be selected, with probability tending to 1. In addition to this algorithm leading to estimators that are consistent and asymptotically normal, they also have the property that when any of the working models contains the true data generating distribution, the resulting estimator attains the semiparametric efficiency bound. Since the estimator corresponding to working model 1 defined in the previous paragraph is the unadjusted estimator, we have that the above method produces an estimator with asymptotic variance as small as or smaller than that of the unadjusted estimator.

A different approach than starting with a fixed list of K working models, is to allow model selection from a large space of possible initial density estimators (not necessarily restricting to the class of generalized linear models), but then using a pre-specified generalized linear model to update this initial density. That is, one could use a machine learning algorithm such as likelihood-based cross-validation (van der Laan and Dudoit, 2004) to produce an initial estimator $p_{01}(Y|A,V)$ of the density of Y given treatment A and

[12]To deal with the possibility that the minimum asymptotic variance corresponding to the list of working models is shared by more than one working model (that is, the possibility of a "tie" for smallest asymptotic variance), we could also add a penalization term, to break any such ties. For example, for each working model k in a list of nested working models, we could add a penalty of $q_k(\log n)/\sqrt{n}$ to the corresponding estimated asymptotic variance (7), where q_k is the number of degrees of freedom in working model k, and select the working model with smallest value of this penalized estimated asymptotic variance.

DOI: 10.2202/1557-4679.1138

baseline variables V (corresponding to step 1 of the targeted maximum likelihood algorithm from Section 6.1). Then steps 2-5 of the targeted maximum likelihood algorithm could be carried out by fitting a pre-specified generalized linear model with canonical link function g, and using $g(E_{p_{01}}(Y|A_i, V_i))$ as an offset; we would require that an intercept and the treatment term A be included as terms in the linear part of the generalized linear model. Proving consistency and asymptotic normality for such a procedure would require verifying the empirical process conditions in Theorem 1 of van der Laan and Rubin (2006), which we list in Appendix 8.1.

Both of the above model selection approaches require pre-specification of the model selection algorithm. Without such pre-specification, there is a general danger in data snooping, e.g. conducting multiple analyses and reporting only the one with largest estimated effect, which can lead to bias and/or inflated p-values. Even when a model selection procedure is pre-specified, one must still prove (as sketched in the previous two paragraphs) that the procedure will lead to a consistent, asymptotically normal estimator.

References

Bickel, P. J. and K. A. Doksum (2001). *Mathematical Statistics*, Volume 1. Upper Saddle River, New Jersey: Prentice Hall.

Bickel, P. J., C. A. Klaassen, Y. Ritov, and J. A. Wellner (1993). *Efficient and Adaptive Estimation for Semiparametric Models*. New York: The Johns Hopkins University Press. Springer-Verlag.

Freedman, D. A. (2008a). On regression adjustments to experimental data. *Advances in Applied Mathematics 40*, 180–193.

Freedman, D. A. (2008b). On regression adjustments to experiments with several treatments. *Annals of Applied Statistics 2*, 176–96.

Freedman, D. A. (2008c). Randomization does not justify logistic regression. *Statistical Science 23*, 237–249.

Gail, M. H. (1986). Adjusting for covariates that have the same distribution in exposed and unexposed cohorts. *In Modern Statistical Methods in Chronic Disease Epidemiology, Eds. S.H. Moolvankar and R. L. Prentice, New York, Wiley.*, 3–18.

Leon, S., A. A. Tsiatis, and M. Davidian (2003). Semiparametric efficient estimation of treatment effect in a pretest-posttest study. *Biometrics 59*, 1046–1055.

Lindsey, J. and P. Lambert (1998). On the appropriateness of marginal models for repeated measurements in clinical trials. *Statistics in Medicine 17*, 447–469.

McCullagh, P. and J. A. Nelder (1998). *Generalized Linear Models* (2nd ed.). Boca Raton, Florida: Chapman and Hall/CRC, Monographs on Statistics and Applied Probability 37.

Moore, K. L. and M. J. van der Laan (2007, April). Covariate adjustment in randomized trials with binary outcomes: Targeted maximum likelihood estimation. *U.C. Berkeley Division of Biostatistics Working Paper Series. Working Paper 215. http://www.bepress.com/ucbbiostat/paper215*.

Neugebauer, R. and M. J. van der Laan. (2002). Why prefer double robust estimates? illustration with causal point treatment studies. working paper 115. http://www.bepress.com/ucbbiostat/paper115. *U.C. Berkeley Division of Biostatistics Working Paper Series*.

Pocock, S. J., S. Assmann, L. Enos, and L. Kasten (2002). Subgroup analysis, covariate adjustment, and baseline comparisons in clinical trial reporting: current practice and problems. *Statistics in Medicine 21*(19), 2917–2930.

Polley, E. and M. van der Laan (2009). "Selecting optimal treatments based on predictive factors". In K. E. Peace (Ed.), *Design and Analysis of Clinical Trials with Time-to-Event Endpoints*, pp. 441–454. Boca Raton: Chapman and Hall/CRC.

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

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

Robins, J. M. (2000). Robust estimation in sequentially ignorable missing data and causal inference models. *Proceedings of the American Statistical Association Section on Bayesian Statistical Science 1999.*, 6–10.

DOI: 10.2202/1557-4679.1138

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

Robinson, L. D. and N. Jewell (1991). Some surprising results about covariate adjustment in logistic regression models. *International Statistical Review 59*, 227–240.

Rosenbaum, P. R. (2002). Covariance adjustment in randomized experiments and observational studies. *Statistical Science 17*(3), 286–327.

Rosenblum, M. and M. van der Laan (2009). Using regression to analyze randomized trials: Valid hypothesis tests despite incorrectly specified models. *Biometrics 65*(3), 937–94.

Rubin, D. B. and M. J. van der Laan (2008). Empirical efficiency maximization: Improved locally efficient covariate adjustment in randomized experiments and survival analysis. *The International Journal of Biostatistics. Available at: http://www.bepress.com/ijb/vol4/iss1/5 4*(1).

Tsiatis, A. A. (2006). *Semiparametric Theory and Missing Data*. Springer Science and Business Media, LLC.

Tsiatis, A. A., M. Davidian, M. Zhang, and X. Lu (2008). Covariate adjustment for two-sample treatment comparisons in randomized clinical trials: A principled yet flexible approach. *Statistics in Medicine 27*, 4658–4677.

van der Laan, M. J. (2010, February). Targeted maximum likelihood based causal inference. *U.C. Berkeley Division of Biostatistics Working Paper Series. Working Paper 259. http://www.bepress.com/ucbbiostat/paper259*.

van der Laan, M. J. and S. Dudoit (2004). Asymptotic optimality of likelihood based cross-validation. *Statistical Applications in Genetics and Molecular Biology 3*(1), 131–154.

van der Laan, M. J. and S. Gruber (2009, April). Collaborative double robust targeted penalized maximum likelihood estimation. *U.C. Berkeley Division of Biostatistics Working Paper Series. Working Paper 246. http://www.bepress.com/ucbbiostat/paper246*.

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

van der Laan, M. J., S. Rose, and S. Gruber (2009). Readings in targeted maximum likelihood estimation. *U.C. Berkeley Division of Biostatistics Working Paper Series. Working Paper 254. http://www.bepress.com/ucbbiostat/paper254*.

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

van der Vaart, A. W. (1998). *Asymptotic Statistics*. New York: Cambridge University Press.

Yang, L. and A. A. Tsiatis (2001, November). Efficiency study of estimators for a treatment effect in a pretest-posttest trial. *The American Statistician 55*(4).

Zhang, M., A. A. Tsiatis, and M. Davidian (2008). Improving efficiency of inferences in randomized clinical trials using auxiliary covariates. *Biometrics 64*(3), 707–715.

DOI: 10.2202/1557-4679.1138