Johns Hopkins University, Dept. of Biostatistics Working Papers

3-3-2011

SIMPLE EXAMPLES OF ESTIMATING CAUSAL EFFECTS USING TARGETED MAXIMUM LIKELIHOOD ESTIMATION

Michael Rosenblum
Johns Hopkins Bloomberg School of Public Health, Department of Biostatistics, mrosenbl@jhsph.edu

Mark J. van der Laan
University of California, Berkeley

Suggested Citation
Rosenblum, Michael and van der Laan, Mark J., "SIMPLE EXAMPLES OF ESTIMATING CAUSAL EFFECTS USING TARGETED MAXIMUM LIKELIHOOD ESTIMATION" (March 2011). *Johns Hopkins University, Dept. of Biostatistics Working Papers*. Working Paper 209.
http://biostats.bepress.com/jhubiostat/paper209

This working paper is hosted by The Berkeley Electronic Press (bepress) and may not be commercially reproduced without the permission of the copyright holder.
Copyright © 2011 by the authors

1 Single Time Point Treatment

We present a brief example, in the context of an observational study of HIV positive individuals on antiretroviral therapy. Assume we have a binary exposure A_0, such as medication adherence being above 90% or not, and a binary outcome Y, such as virologic failure. Assume we have baseline variables L_0 that should include all important confounders of the effect of A_0 on Y.

Say we want to estimate the causal effect of A_0 on the mean of Y, as a risk difference; that is, we'd like to estimate the difference between the population mean of Y were everyone to have had exposure set as $A_0 = 0$, and the population mean were everyone to have had exposure set as $A_0 = 1$. (Below, we use both the terms "exposure" and "treatment" to refer to A_0.) Below, for simplicity, we just show how to estimate the treatment specific (counterfactual) mean setting $A_0 = 1$.

Let p denote a joint probability density on the variables (L_0, A_0, Y). (Throughout, we use the term "density" in the general sense; that is, it refers to a frequency function for discrete valued variables and refers to a density for continuous valued variables.) Here we will put no restrictions on p, except that we only consider p for which all the conditional distributions we give below are well-defined. Assume that for each subject i we get a vector of data $(L_0^{(i)}, A_0^{(i)}, Y^{(i)})$, where each such vector is an independent draw from the true (unknown) data generating distribution p^* on (L_0, A_0, Y). Assume we have n subjects.

Under certain assumptions, the treatment specific mean of Y setting $A_0 = 1$ equals the mean over the baseline variables L_0 of $p^*(Y = 1|A_0 = 1, L_0)$, which we denote by

$$\psi_1(p^*) = \sum_{l_0} p^*(Y = 1|A_0 = 1, L_0 = l_0)p^*(L_0 = l_0). \tag{1}$$

If the distribution of L_0 is continuous, the above sum would be replaced by an integral. The one-dimensional parameter we will estimate in this section is (1), in the nonparametric model. By "one-dimensional parameter in the nonparametric model" we mean a function from the space of all distributions p on the variables (L_0, A_0, Y) to the real line.[1] As described in (van der Laan, 2006), the parameter (1) may be of interest even when the assumptions that allow interpretation of it as a treatment specific mean do not hold.

We will estimate (1) by first getting a suitable estimate for $p^*(Y = 1|A_0 = 1, L_0)$, and then averaging it over the marginal distribution of L_0 that we have in the data (that is, averaging over the empirical distribution of L_0). This is a substitution estimator for the parameter (1), and therefore has the desirable property that the estimated parameter value is in the range of possible parameter values; here that means our estimate of (1) will always be in the interval $[0, 1]$. Targeted maximum likelihood gives a way to estimate $p^*(Y = 1|A_0 = 1, L_0)$ that is, roughly speaking, targeted at minimizing the mean squared error of the parameter (1) we're interested in.

In our case, targeted maximum likelihood estimation involves the following steps:

a. obtaining an initial estimate $\hat{p}$ for the joint density of (L_0, A_0, Y);

[1] More precisely, we consider all distributions p on the variables (L_0, A_0, Y) that have a density with respect to a dominating measure μ.

CODRA
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

b. constructing a "least-favorable" parametric model $\{p_\epsilon\}$ for the parameter (1) at the density $\hat{p}$;

c. fitting this parametric model with maximum likelihood estimation to obtain an updated density p';

d. estimating the parameter (1) using the substitution estimator $\psi_1(p')$ (that is, evaluating the expression (1) using the estimated density p' in place of the true (unknown) density p^*).

The general targeted maximum likelihood algorithm involves iterating steps (b) and (c) above until convergence; however, in the cases we consider in this paper, a single iteration suffices.

An important step in targeted maximum likelihood estimation is the construction of a "least favorable" parametric model for the parameter of interest. Formally, such a parametric model $\{p_\epsilon\}$ needs to satisfy two conditions: (i) it must equal the current density estimate $\hat{p}$ at $\epsilon = 0$, and (ii) the linear span of its score at $\epsilon = 0$ must contain the efficient influence function at the current density estimate. The efficient influence function for the parameter (1) in the nonparametric model, at density p, as given for example in (Scharfstein et al., 1999; Rosenblum and van der Laan, 2010a,b), is

$$\left\{ \frac{A_0}{p(A_0 = 1 \mid L_0)} \left[Y - p(Y = 1 \mid A_0 = 1, L_0) \right] \right\} + \left\{ p(Y = 1 \mid A_0 = 1, L_0) - \psi_1(p) \right\}. \tag{2}$$

Intuitively, the efficient influence function gives the direction in which the parameter we are estimating is most sensitive, to first order (when p equals the true data generating distribution). Below, we show how to use logistic regression models to construct such least favorable models satisfying conditions (i) and (ii).

We now present one possible targeted maximum likelihood estimator for the parameter (1). Step (a) is to specify initial estimators for the true (unknown) data generating distribution p^*. In general, nonparametric methods such as kernel smoothing or machine learning algorithms that use cross-validation, can be used to construct such estimators. For clarity of exposition, here we use parametric model fits as our initial estimators. We fit an initial logistic regression of Y on A_0 and L_0, such as

$$p_\alpha(Y = 1 | A_0, L_0) = \text{expit}(\alpha_0 + \alpha_1 A_0 + \alpha_2 L_0). \tag{3}$$

Any terms that are functions of A_0 and/or L_0 can be included in the model. Next, we fit an initial logistic regression of A_0 on L_0, such as

$$p_\beta(A_0 = 1 | L_0) = \text{expit}(\beta_0 + \beta_1 L_0 + \beta_1 L_0^2). \tag{4}$$

Any terms that are functions L_0 can be included in the model.

Denote the estimated coefficients from fitting the above logistic regression models by $\hat{\alpha}$ and $\hat{\beta}$. We denote the model fit for $p(A_0 = 1 | L_0)$ by

$$\hat{p}(A_0 = 1 | L_0) := \text{expit}(\hat{\beta}_0 + \hat{\beta}_1 L_0 + \hat{\beta}_1 L_0^2),$$

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/jhubiostat/paper209

and analogously define $\hat{p}(Y = 1|A_0, L_0)$. Let $\hat{p}(L_0)$ denote the empirical distribution of L_0, which gives mass $1/n$ to each observation $L_0^{(i)}$. Our initial estimator of the joint density of (L_0, A_0, Y) is $\hat{p} := \hat{p}(Y|A_0, L_0)\hat{p}(A_0|L_0)\hat{p}(L_0)$.

For the above choice of initial estimators, the targeted maximum likelihood estimator we present below is identical to the estimator given in the Rejoinder to Comments in Scharfstein et al. (1999) on page 1141. For more information on the relationship between estimators here and prior work, we refer the reader to Appendix 2 of (Rosenblum and van der Laan, 2010b).

We now turn to step (b) of the targeted maximum likelihood algorithm. This involves constructing a least favorable parametric model $\{p_\epsilon\}$ satisfying criteria (i) and (ii) above. Our parametric model keeps the components $\hat{p}(A_0|L_0)$ and $\hat{p}(L_0)$ of the initial density estimate fixed; we only perturb the component $\hat{p}(Y|A_0, L_0)$ using the logistic regression model

$$p_\epsilon(Y = 1|A_0, L_0) = \text{expit}(\epsilon C(A_0, L_0) + \hat{\alpha}_0 + \hat{\alpha}_1 A_0 + \hat{\alpha}_2 L_0), \tag{5}$$

where the $\hat{\alpha}_i$ are considered fixed and $C(A_0, L_0)$ is a "clever covariate" that we define next.

The clever covariate $C(A_0, L_0)$ is chosen so that condition (ii) given above is satisfied for the above logistic regression model. (Condition (i) is automatically satisfied, since at $\epsilon = 0$, we have (5) equals $\hat{p}(Y = 1|A_0, L_0)$.) Condition (ii) states that the linear span of the score of the logistic regression model at $\epsilon = 0$ must contain the efficient influence function (2). As argued[2] in (Moore and van der Laan, 2007; Rosenblum and van der Laan, 2010b), since we chose the initial distribution $\hat{p}(L_0)$ to be the empirical distribution of L_0, it suffices that we define $C(A_0, L_0)$ so that the score of the above logistic regression model at $\epsilon = 0$ equals the term in braces on the left in (2) evaluated at $p = \hat{p}$, which we reproduce below:

$$\frac{A_0}{\hat{p}(A_0 = 1 \mid L_0)} \left[Y - \hat{p}(Y = 1 \mid A_0 = 1, L_0)\right]. \tag{7}$$

The score of the logistic regression model (5) at $\epsilon = 0$ (as derived in the Appendix) is

$$C(A_0, L_0) \left[Y - \hat{p}(Y = 1|A_0, L_0)\right]. \tag{8}$$

This motives defining the clever covariate to be

$$C(A_0, L_0) := A_0/\hat{p}(A_0 = 1|L_0),$$

[2]To be more formal, we would construct a least favorable parametric model satisfying (i) and (ii) that has two parameters ϵ and τ, and that fluctuates $\hat{p}(Y|A_0, L_0)$ as in the logistic regression model (5), and fluctuates the marginal density $\hat{p}(L_0)$ using the following one dimensional model:

$$p_\tau(L_0) := s_\tau \exp(\tau \left[\hat{p}(Y = 1 \mid A_0 = 1, L_0) - \psi_1(\hat{p})\right])\hat{p}(L_0), \tag{6}$$

where the constant $s_\tau := 1/[\frac{1}{n}\sum_{i=1}^n \exp\left(\tau\left[\hat{p}(Y = 1 \mid A_0 = 1, L_0^{(i)}) - \psi_1(\hat{p})\right]\right)$ is chosen so that $p_\tau(l_0)$ sums to 1 for each τ. The score of the above model at $\tau = 0$ equals the term in braces on the right in (2) at $p = \hat{p}$. Thus the linear span of this score and the score of (5) at $\epsilon = 0$ (given in (7) below), contains the efficient influence function (2) at $\hat{p}$, as required in condition (ii) above. We show in Section 3 of (Rosenblum and van der Laan, 2010b) that the maximum likelihood estimator $\hat{\tau}$ of τ is always 0 (which can be deduced from $\hat{p}(L_0)$ being the empirical distribution of L_0). Thus, there is no update to the density $\hat{p}(L_0)$, and the the resulting targeted maximum likelihood estimator is the same as given in the text where the component (6) of the parametric model is ignored. We emphasize this relied on choosing $\hat{p}(L_0)$ to be the empirical distribution of L_0.

Research Archive

Hosted by The Berkeley Electronic Press

which ensures that the score (8) equals (7), thereby satisfying condition (ii) above.

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; Rosenblum and van der Laan, 2010a,b).

We now update our estimate of $p^*(Y = 1|A_0, L_0)$, by fitting the logistic regression model (5) where the $\hat{\alpha}_i$ are considered fixed (they were computed above in (3)) and the only variable is ϵ. This can be done by entering $\hat{\alpha}_0 + \hat{\alpha}_1 A_0 + \hat{\alpha}_2 L_0$ as an offset in the logistic regression. Fitting this logistic regression model gives an estimate $\hat{\epsilon}$ for ϵ. Our final estimate for $p^*(Y = 1|A_0, L_0)$ is then

$$\text{expit}(\hat{\epsilon} C(A_0, L_0) + \hat{\alpha}_0 + \hat{\alpha}_1 A_0 + \hat{\alpha}_2 L_0),$$

and from this we get that our final estimate for $p^*(Y = 1|A_0 = 1, L_0)$ is

$$\text{expit}(\hat{\epsilon} C(1, L_0) + \hat{\alpha}_0 + \hat{\alpha}_1 + \hat{\alpha}_2 L_0). \tag{9}$$

Our estimate for the parameter (1) is the substitution estimator using (9), that is, the average over the empirical distribution of L_0 of (9), which is

$$\frac{1}{n} \sum_{i=1}^{n} \text{expit}(\hat{\epsilon} C(1, L_0^{(i)}) + \hat{\alpha}_0 + \hat{\alpha}_1 + \hat{\alpha}_2 L_0^{(i)}).$$

where $L_0^{(i)}$ is the value of L_0 from the ith subject. Under regularity conditions, this estimator is a doubly robust, locally efficient estimator of ψ_1, which means that if at least one of the models (3) or (4) is correctly specified, then the above estimator is consistent and asymptotically normal; if both models are correctly specified it is also efficient. For this particular choice of initial estimators, this was shown in the Rejoinder to Comments in Scharfstein et al. (1999) on page 1141.

For an extension of the above construction to outcomes that are not binary, e.g. for Y continuous or a nonnegative integer, see examples in e.g. (Rosenblum and van der Laan, 2010a). There, the same methods as above are given, but replacing logistic regression by e.g. Poisson regression for count data.

Here is R code to compute the above estimator:

```
# Given outcomes Y, treatment A, baseline variables L,
# all of length n:

# 1. Fit initial models (2) and (3) from text:
initial_model_for_Y <- glm(Y ~ 1 + A + L, family=binomial)
initial_model_for_A  <- glm(A ~ 1 +L + L^2, family=binomial)

# 2. Compute clever covariate:
clever_covariate <-
  A/predict.glm(initial_model_for_A,type="response")
# Create offset:
offset_vals <- predict.glm(initial_model_for_Y)
```


http://biostats.bepress.com/jhubiostat/paper209

```
# 3. Refit model for Y given A, L, with clever cov. and offset:
updated_model_for_Y <-
  glm(Y ~ clever_covariate-1, family=binomial,offset=offset_vals)

# 4. Compute final estimate (6) from text:
# First compute clever covariate setting A to 1:
expit <- function(x){return(exp(x)/(1+exp(x)))}
clever_covariate_setting_A_to_1 <-
  1/predict.glm(initial_model_for_A,type="response")
final_estimate<- mean(expit(
  updated_model_for_Y$coefficients*clever_covariate_setting_A_to_1
  + initial_model_for_Y$coefficients %*% rbind(1,rep(1,n),L)))
```

2 Time Dependent Treatments

We now consider a case where we have two time points of treatment, and we want to estimate the mean outcome that would result from setting the treatment at both time points. It is straightforward to generalize these methods to dynamic treatments (that is, where treatment is a function of prior measurements and/or treatments). It is also straightforward to extend this to deal with missing data.

A_0, A_1 are the treatments, e.g. type of antiretroviral regimen at times 0 and 1. L_0, L_1 are measurements such as CD4 count, viral load, etc. that occur before each treatment. We let Y be the final outcome (death or not). The list of variables, in the order they are measured on each subject, are: L_0, A_0, L_1, A_1, Y, where L_0 are baseline variables; A_0 is type of first antiretroviral regimen; L_1 is a set of measurements made after A_0, such as viral load, death or not, CD4, etc.; A_1 is regimen at next time point; Y is death or not at the following time point.

Consider estimating the mean outcome Y under the treatment strategy of setting everyone to have $A_0 = a_0$ and $A_1 = a_1$. That is, we want to know what the probability of death would be, had everyone been assigned to antiretroviral therapy a_0 at time 0 and antiretroviral therapy a_1 at time 1. We could then compare, say, the effect of setting $a_0 = a_1 =$ PI therapy vs. $a_0 = a_1 =$ NNRTI therapy. (The same methods can be generalized to estimate the effect of dynamic treatments, of the form: if L_0 is larger than some threshold c, then assign treatment a_0, else assign a different treatment.)

We assume here that all variables except L_0 are binary, for simplicity. Extensions to the more useful case of L_1 being e.g. multivariate containing continuous and/or categorical variables are given in (van der Laan, 2010a,b).

Under certain assumptions, the counterfactual mean of Y setting $A_0 = a_0$ and $A_1 = a_1$ is equal to the g-computation formula of Robins:

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

$$\psi_2^* := \sum_{l_0} \sum_{l_1} p^*(Y = 1 \mid A_1 = a_1, L_1 = l_1, A_0 = a_0, L_0 = l_0) \times$$
$$p^*(L_1 = l_1 \mid A_0 = a_0, L_0 = l_0) p^*(L_0 = l_0), \tag{10}$$

where p^* is the true (unknown) density of the variables. This is the two time point analog to the formula (1) above for a single time point treatment. We give the efficient influence function for the above parameter, in the nonparametric model, in the Appendix.

In what follows, we estimate the value of the above display at $a_0 = a_1 = 1$, that is, the counterfactual mean of Y setting treatments A_0, A_1 to 1. Estimating this at other values of a_0, a_1 is similar. Roughly speaking, targeted maximum likelihood estimation finds estimates for each of the conditional probabilities in the above formula, in a way targeted at estimating the overall parameter (10).

For each subject i, assume we have a vector of data $(L_0^{(i)}, A_0^{(i)}, L_1^{(i)}, A_1^{(i)}, Y^{(i)})$. Assume each such vector is an independent draw from an unknown density (or frequency function) p^* on (L_0, A_0, L_1, A_1, Y). We put no constraints on p^* except that the conditional distributions we give below are well-defined.

Here is one example of a targeted maximum likelihood estimator for our problem. As in the single time point case, the first step is to specify initial estimators for the true data generating distribution p^*. We emphasize that in general, nonparametric methods can be used to construct such estimators. For clarity of exposition, here we use parametric model fits as our initial estimators.

We fit the following logistic regression models:

$$p_\beta(Y = 1 \mid A_1, L_1, A_0, L_0) = \text{logit}^{-1}(\beta_0 + \beta_1 L_0 + \beta_2 A_0 + \beta_3 L_1 + \beta_4 A_1), \tag{11}$$
$$p_\alpha(A_1 = 1 | L_1, A_0, L_0) = \text{logit}^{-1}(\alpha_0 + \alpha_1 L_0 + \alpha_2 A_0 + \alpha_3 L_1), \tag{12}$$
$$p_\gamma(L_1 = 1 \mid A_0, L_0) = \text{logit}^{-1}(\gamma_0 + \gamma_1 L_0 + \gamma_2 A_0 + \gamma_3 L_0 A_0), \tag{13}$$
$$p_\tau(A_0 = 1 \mid L_0) = \text{logit}^{-1}(\tau_0 + \tau_1 L_0), \tag{14}$$

We denote the model fits, which we refer to as "initial fits" by $\hat{p}$, e.g. $\hat{p}(A_0 = 1 \mid L_0 = l_0) = \text{logit}^{-1}(\hat{\tau}_0 + \hat{\tau}_1 l_0)$. We let the initial fit $\hat{p}(L_0)$ be the empirical distribution of L_0.

We next define logistic regression models that use "clever covariates," in a manner analogous to the previous section. These clever covariates are chosen to ensure condition (ii) given above holds, for the efficient influence function given in the Appendix. We explain in the appendix how these clever covariates were selected.

Define the following "clever covariate":

$$C_1(l_0', a_0', l_1', a_1') := \frac{1[a_1' = 1] 1[a_0' = 1]}{\hat{p}(A_1 = 1 | L_1 = l_1', A_0 = 1, L_0 = l_0') \hat{p}(A_0 = 1 \mid L_0 = l_0')}, \tag{15}$$

where $1[S]$ is the indicator variable that S is true (so is equal to 1 when S is true, and 0 when it is false). For each subject i, (with data $(L_0^{(i)}, A_0^{(i)}, L_1^{(i)}, A_1^{(i)}, Y^{(i)})$), compute the value of the clever covariate $C_1^{(i)} := C_1(L_0^{(i)}, A_0^{(i)}, L_1^{(i)}, A_1^{(i)})$. Now do a logistic regression of Y on the clever

COBRA
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/jhubiostat/paper209

covariate C_1, using the initial fit $\hat{\beta}_0 + \hat{\beta}_1 L_0 + \hat{\beta}_2 A_0 + \hat{\beta}_3 L_1 + \hat{\beta}_4 A_1$ as offset. That is, fit the logistic regression model

$$\begin{aligned} &p_{\epsilon_1}(Y = 1 \mid A_1, L_1, A_0, L_0) = \\ &\quad \text{logit}^{-1}(\epsilon_1 C_1(L_0, A_0, L_1, A_1) + \hat{\beta}_0 + \hat{\beta}_1 L_0 + \hat{\beta}_2 A_0 + \hat{\beta}_3 L_1 + \hat{\beta}_4 A_1), \end{aligned} \tag{16}$$

where the $\hat{\beta}$ are considered fixed numbers, and the only variable is ϵ_1. Let $\hat{\epsilon}_1$ denote the maximum likelihood estimate of ϵ_1. We now define

$$\begin{aligned} &\hat{p}_{\hat{\epsilon}_1}(Y = 1 \mid A_1, L_1, A_0, L_0) := \\ &\quad \text{logit}^{-1}(\hat{\epsilon}_1 C_1(L_0, A_0, L_1, A_1) + \hat{\beta}_0 + \hat{\beta}_1 L_0 + \hat{\beta}_2 A_0 + \hat{\beta}_3 L_1 + \hat{\beta}_4 A_1). \end{aligned} \tag{17}$$

Next, define another clever covariate,

$$\begin{aligned} C_2(l_0', a_0') \quad := \frac{1[a_0' = 1]}{\hat{p}(A_0 = 1 \mid L_0 = l_0')} \times \\ \{\hat{p}_{\hat{\epsilon}_1}(Y = 1 \mid A_1 = 1, L_1 = 1, A_0 = 1, L_0 = l_0') \\ -\hat{p}_{\hat{\epsilon}_1}(Y = 1 \mid A_1 = 1, L_1 = 0, A_0 = 1, L_0 = l_0')\}. \end{aligned} \tag{18}$$

For each subject i, compute the value of the clever covariate $C_2^{(i)} := C_2(L_0^{(i)}, A_0^{(i)})$. Now do a logistic regression of L_1 on the clever covariate C_2, using the initial fit $\hat{\gamma}_0 + \hat{\gamma}_1 L_0 + \hat{\gamma}_2 A_0 + \hat{\gamma}_3 L_0 A_0$ as offset. That is, fit the logistic regression model

$$p_{\epsilon_2}(L_1 = 1 \mid A_0, L_0) = \text{logit}^{-1}(\epsilon_2 C_2(L_0, A_0) + \hat{\gamma}_0 + \hat{\gamma}_1 L_0 + \hat{\gamma}_2 A_0 + \hat{\gamma}_3 L_0 A_0), \tag{19}$$

where the $\hat{\gamma}$ are considered fixed numbers, and the only variable is ϵ_2. Let $\hat{\epsilon}_2$ denote the maximum likelihood estimate of ϵ_2. We now define

$$\hat{p}_{\hat{\epsilon}_2}(L_1 = 1 \mid A_0, L_0) := \text{logit}^{-1}(\hat{\epsilon}_2 C_2(L_0, A_0) + \hat{\gamma}_0 + \hat{\gamma}_1 L_0 + \hat{\gamma}_2 A_0 + \hat{\gamma}_3 L_0 A_0). \tag{20}$$

Lastly, we compute the substitution estimator for (10) at the above model fits. That is, we evaluate (10) at $a_0 = a_1 = 1$ by substituting estimated densities (17) and (20) for the corresponding true densities, and using the empirical distribution for L_0 (which we denote by $\hat{p}(L_0 = l_0)$. That is, our final estimate of the mean of Y setting A_0, A_1 both equal to 1, is

$$\begin{aligned} &\sum_{l_0 \in \mathcal{L}_0} \sum_{l_1 \in \{0,1\}} \hat{p}_{\hat{\epsilon}_1}(Y = 1 \mid A_1 = 1, L_1 = l_1, A_0 = 1, L_0 = l_0) \times \\ &\qquad \hat{p}_{\hat{\epsilon}_2}(L_1 = l_1 \mid A_0 = 1, L_0 = l_0)\hat{p}(L_0 = l_0). \\ &= \frac{1}{n} \sum_{i=1}^{n} \sum_{l_1 \in \{0,1\}} \hat{p}_{\hat{\epsilon}_1}(Y = 1 \mid A_1 = 1, L_1 = l_1, A_0 = 1, L_0 = L_0^{(i)}) \times \\ &\qquad \hat{p}_{\hat{\epsilon}_2}(L_1 = l_1 \mid A_0 = 1, L_0 = L_0^{(i)}), \end{aligned}$$

where $\mathcal{L}_0$ denotes the set of possible values L_0 can take.

Under regularity conditions, this estimator is doubly robust, locally efficient; this means that if the models (11) and (13) are correctly specified, or if the models (12) and (14) are correctly specified, then the above estimator is consistent and asymptotically normal; if all four models are correctly specified it is also efficient.

SCDHA
Scholarly Commons
Research Archive

Hosted by The Berkeley Electronic Press

3 Appendix

In the first part of the Appendix, we derive the score of the logistic regression models used above. In the second part of the Appendix, we present the efficient influence function for the parameter (10) in the nonparametric model from Section 2.

3.1 Score of Logistic Regression Models

Consider the logistic regression model (5) from Section 1. We derive its score. The score of a parametric model, by definition, is the derivative of the log-likelihood. We only consider the component of the likelihood that depends on the parameter ϵ, since the components that do not change with ϵ have derivative 0 and make no contribution to the score. In Section 1, for a single observation (L_0, A_0, Y), the likelihood of outcome Y given L_0, A_0 under the logistic regression model (5) is $p_\epsilon(Y = 1|A_0, L_0)$ when $Y = 1$ and is $1 - p_\epsilon(Y = 1|A_0, L_0)$ when $Y = 0$. This conditional likelihood can be expressed, equivalently, as

$$L(\epsilon; L_0, A_0, Y) = p_\epsilon(Y = 1|A_0, L_0)^Y (1 - p_\epsilon(Y = 1|A_0, L_0))^{1-Y}.$$

The derivative of the log of this likelihood is

$$\begin{aligned}
\frac{d}{d\epsilon}\left[\log L(\epsilon; L_0, A_0, Y)\right] &= Y\frac{\frac{d}{d\epsilon}p_\epsilon(Y = 1|A_0, L_0)}{p_\epsilon(Y = 1|A_0, L_0)} + (1 - Y)\frac{-\frac{d}{d\epsilon}p_\epsilon(Y = 1|A_0, L_0)}{1 - p_\epsilon(Y = 1|A_0, L_0)} \\
&= \left[\frac{Y}{p_\epsilon(Y = 1|A_0, L_0)} - \frac{1 - Y}{1 - p_\epsilon(Y = 1|A_0, L_0)}\right]\frac{d}{d\epsilon}p_\epsilon(Y = 1|A_0, L_0) \\
&= \left[Y(1 - p_\epsilon(Y = 1|A_0, L_0)) - (1 - Y)(p_\epsilon(Y = 1|A_0, L_0))\right] C(A_0, L_0) \\
&= \left[Y - p_\epsilon(Y = 1|A_0, L_0)\right] C(A_0, L_0), \qquad (21)
\end{aligned}$$

where the third equality follows from the following property of our logistic regression model:

$$\frac{d}{d\epsilon}p_\epsilon(Y = 1|A_0, L_0) \quad = \quad p_\epsilon(Y = 1|A_0, L_0)(1 - p_\epsilon(Y = 1|A_0, L_0))C(A_0, L_0).$$

The score at $\epsilon = 0$ is (21) setting $\epsilon = 0$, which (because at $\epsilon = 0$ we have $p_\epsilon(Y = 1|A_0, L_0) = \hat{p}(Y = 1|A_0, L_0)$ from Section 1 by construction) is identical to (8), as desired.

The scores of the logistic regression models (16) and (19) in Section 2 are obtained similarly.

3.2 Efficient Influence Function for the Parameter (10) in the Nonparametric Model

Consider the parameter defined in (10), which we reproduce here:

$$\begin{aligned}
\psi(p^*) := \sum_{l_0}\sum_{l_1} & p^*(Y = 1 \mid A_1 = a_1, L_1 = l_1, A_0 = a_0, L_0 = l_0) \times \\
& p^*(L_1 = l_1 \mid A_0 = a_0, L_0 = l_0)p^*(L_0 = l_0), \qquad (22)
\end{aligned}$$

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/jhubiostat/paper209

where p^* is the true (unknown) density of the variables. This is the two time point analog to the formula (1) above for a single time point treatment. Below we give the efficient influence function for this parameter. In Theorem 1 of (van der Laan, 2010a), the efficient influence function for this parameter, and for a more general set of parameters (including parameters defined by dynamic regimes), is provided.

Recall we made the simplifying assumption that L_1 is binary valued. Denote the inner summation in (22) (and leaving out the term $p(L_0 = l_0)$) by

$$\begin{aligned}
&f(p, a_1, a_0, l_0) \\
&= \sum_{l_1} p(Y = 1 \mid A_1 = a_1, L_1 = l_1, A_0 = a_0, L_0 = l_0) \times p(L_1 = l_1 \mid A_0 = a_0, L_0 = l_0) \\
&= p(Y = 1 \mid A_1 = a_1, L_1 = 1, A_0 = a_0, L_0 = l_0) \times p(L_1 = 1 \mid A_0 = a_0, L_0 = l_0) \\
&\quad + p(Y = 1 \mid A_1 = a_1, L_1 = 0, A_0 = a_0, L_0 = l_0) \times p(L_1 = 0 \mid A_0 = a_0, L_0 = l_0) \quad (23)
\end{aligned}$$

The efficient influence function for the above parameter (22) for any treatments a_0, a_1, in the nonparametric model at density p is:

$$D(p, L_0, A_0, L_1, A_1, Y) = D_0(p, L_0) + D_1(p, L_0, A_0, L_1) + D_2(p, L_0, A_0, L_1, A_1, Y) \quad (24)$$

where

$$D_0(p, L_0) = f(p, a_1, a_0, L_0) - \psi(p), \quad (25)$$

and

$$\begin{aligned}
D_1(p, L_0, A_0, L_1) &= \frac{1[A_0 = a_0]}{p(A_0 = a_0 \mid L_0)} \\
&\times [p(Y = 1|A_1 = a_1, L_1 = 1, A_0 = a_0, L_0) - p(Y = 1|A_1 = a_1, L_1 = 0, A_0 = a_0, L_0)] \\
&\times [L_1 - p(L_1 = 1 \mid A_0 = a_0, L_0)], \quad (26)
\end{aligned}$$

and

$$\begin{aligned}
&D_2(p, L_0, A_0, L_1, A_1, Y) \\
&= \frac{1[A_1 = a_1]1[A_0 = a_0]}{p(A_1 = a_1|L_1, A_0 = a_0, L_0)p(A_0 = a_0 \mid L_0)} \{Y - p(Y = 1|A_1 = a_1, L_1, A_0 = a_0, L_0)\}. \quad (27)
\end{aligned}$$

We suppress the dependence of the above functions D, D_0, D_1, and D_2 on a_0, a_1 for clarity. We show at the end of this section that the component D_1 of the efficient influence function can be equivalently expressed in another form, which may be more familiar to the reader.

We now consider how the clever covariates C_1 and C_2 from Section 2 were chosen, based on the above representation of the efficient influence function (24). Recall that our goal in constructing clever covariates is to ensure the linear span of the scores of the logistic regression models (16) and (19) at $\epsilon = (\epsilon_1, \epsilon_2) = \mathbf{0}$ contains the efficient influence function (24) at the initial fit $p = \hat{p}$. Also, recall that we set $a_1 = a_0 = 1$, for simplicity, in Section 2; we consider these values of a_0, a_1 in what follows.

COBRA
BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

The score of logistic regression model (16) at $\epsilon = \mathbf{0}$ is

$$C_1(L_0, A_0, L_1, A_1) \left\{ Y - \text{logit}^{-1}(\hat{\beta}_0 + \hat{\beta}_1 L_0 + \hat{\beta}_2 A_0 + \hat{\beta}_3 L_1 + \hat{\beta}_4 A_1) \right\}.$$

Notice the similarity in form to the component (27) of the efficient influence function (24). We chose $C_1(L_0, A_0, L_1, A_1)$ in (15) above so that this score equals the component (27) of the efficient influence function at the initial fit $p = \hat{p}$.

The score of logistic regression model (19) at $\epsilon = \mathbf{0}$ is

$$C_2(L_0, A_0) \left\{ L_1 - \text{logit}^{-1}(\gamma_0 + \gamma_1 L_0 + \gamma_2 A_0 + \gamma_3 L_0 A_0) \right\}.$$

Notice the similarity in form to the component (26) of the efficient influence function. We chose $C_2(L_0, A_0)$ in (18) above so that this score equals the component (26) of the efficient influence function at the initial fit $p = \hat{p}$.

As argued in (Moore and van der Laan, 2007) (and in footnote 2 in Section 1 of this paper for the single time point case) as long as our initial estimator for the marginal distribution of L_0 is the corresponding empirical distribution, it is not necessary to fit an additional model whose score equals (25).

Using the clever covariates defined in (15) and (18), we then have (with the caveat in the previous paragraph which allows us to exclude the component (25)) that the efficient influence function given above is in the linear span of the scores of the models (16) and (19) at $\epsilon = \mathbf{0}$, so that condition (ii) is satisfied.

We point out that the component $D_1(p, L_0, A_0, L_1)$, defined in (26), of the efficient influence function (24) can be equivalently written in the more familiar form:

$$\frac{1[A_0 = a_0]}{p(A_0 = 1 \mid L_0)} \left\{ p(Y = 1 | A_1 = a_1, L_1, A_0 = a_0, L_0) - f(p, a_1, a_0, L_0) \right\}, \tag{28}$$

To show this, first note that because L_1 is binary-valued, we have

$$\begin{aligned} p(Y = 1 | A_1 = a_1, L_1, A_0 = a_0, L_0) &= L_1 p(Y = 1 | A_1 = a_1, L_1 = 1, A_0 = a_0, L_0) \\ &\quad + (1 - L_1) p(Y = 1 | A_1 = a_1, L_1 = 0, A_0 = a_0, L_0). \end{aligned}$$

Then by (23), we have that the expression in braces in (28) equals

$$\begin{aligned} & p(Y = 1 | A_1 = a_1, L_1, A_0 = a_0, L_0) - f(p, a_1, a_0, L_0) \\ &= p(Y = 1 | A_1 = a_1, L_1 = 1, A_0 = a_0, L_0) \left[L_1 - p(L_1 = 1 \mid A_0 = a_0, L_0 = l_0) \right] \\ &\quad + p(Y = 1 | A_1 = a_1, L_1 = 0, A_0 = a_0, L_0) \left[(1 - L_1) - p(L_1 = 0 \mid A_0 = a_0, L_0 = l_0) \right] \\ &= \left[p(Y = 1 | A_1 = a_1, L_1 = 1, A_0 = a_0, L_0) - p(Y = 1 | A_1 = a_1, L_1 = 0, A_0 = a_0, L_0) \right] \\ &\quad \times \left[L_1 - p(L_1 = 1 \mid A_0 = a_0, L_0 = l_0) \right]. \end{aligned}$$

This implies (28) equals (26), as desired.

COBRA
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/jhubiostat/paper209

References

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*.

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.

Rosenblum, M. and M. van der Laan (2010a). Simple, Efficient Estimators of Treatment Effects in Randomized Trials Using Generalized Linear Models to Leverage Baseline Variables. *The International Journal of Biostatistics. Article 13. DOI: 10.2202/1557-4679.1138 Available at: http://www.bepress.com/ijb/vol6/iss1/13 6*(1).

Rosenblum, M. and M. J. van der Laan (2010b). Targeted Maximum Likelihood Estimation of the Parameter of a Marginal Structural Model. *The International Journal of Biostatistics 6*(2), http://www.bepress.com/ijb/vol6/iss2/19.

Scharfstein, D. O., A. Rotnitzky, and J. M. Robins (1999). Adjusting for Non-Ignorable Drop-out Using Semiparametric Nonresponse Models, (with Discussion and Rejoinder). *Journal of the American Statistical Association 94*, 1096–1120 (1121–1146).

van der Laan, M. (2006). Statistical Inference for Variable Importance. *International Journal of Biostatistics 2*(1).

van der Laan, M. J. (2010a). Targeted Maximum Likelihood Based Causal Inference: Part I. *The International Journal of Biostatistics. Article 2. DOI: 10.2202/1557-4679.1211 Available at: http://www.bepress.com/ijb/vol6/iss2/2 6*(2).

van der Laan, M. J. (2010b). Targeted Maximum Likelihood Based Causal Inference: Part II. *The International Journal of Biostatistics. Article 3. DOI: 10.2202/1557-4679.1241 Available at: http://www.bepress.com/ijb/vol6/iss2/3 6*(2).

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).

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press