Estimating Causal Effects Using Targeted Maximum Likelihood Estimation

Michael Rosenblum and Mark J. van der Laan

March 16, 2010

Here is brief overview of targeted maximum likelihood for estimating the causal effect of a single time point treatment and of a two time point treatment. We include R code for the single time point case. We present simple examples demonstrating how to apply the methodology developed in (van der Laan and Rubin, 2006; Moore and van der Laan, 2007; van der Laan, 2010a,b).

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, that is, we'd like to estimate what the population mean of Y would be were everyone to have had exposure $A_0 = 0$, and also what the population mean of Y would be were everyone to have had exposure $A_0 = 1$. (Below, we use both the terms "exposure" and "treatment" to refer to A_0.) Below, for simplicity, we just show how to estimate the effect of setting $A_0 = 1$.

Let p denote the density of the true (unknown) data generating distribution. Under certain assumptions, this causal effect equals the mean over the baseline variables L_0 of $p(Y = 1|A_0 = 1, L_0)$, that is,

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

If the distribution of L_0 is continuous, the above sum would be replaced by an integral. The parameter we will estimate in this section is (1). We will estimate it 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 (the empirical mean). Targeted maximum likelihood gives a way to estimate $p(Y = 1|A_0 = 1, L_0)$ that is targeted at minimizing the mean squared error of the parameter (1) we're interested in.

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 an unknown density (or frequency function) $p(L_0, A_0, Y)$. Assume we have n subjects.

We now present one possible targeted maximum likelihood estimator. First, we fit an initial logistic regression of Y on A_0 and L_0, such as

$$p(Y = 1|A_0, L_0) = \text{expit}(\alpha_0 + \alpha_1 A_0 + \alpha_2 L_0). \tag{2}$$

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

$$p(A_0 = 1|L_0) = \text{expit}(\beta_0 + \beta_1 L_0 + \beta_1 L_0^2). \tag{3}$$

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

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

and analogously define $\hat{p}(Y = 1|A_0, L_0)$.

We then compute, for each subject, the value of a "clever covariate," which we will use to update the above initial logistic regression estimate $\hat{p}(Y = 1|A_0, L_0)$. That is, we will define a clever covariate $C(A_0, L_0)$, and update our estimate of $p(Y = 1|A_0, L_0)$, by fitting the logistic regression:

$$p(Y = 1|A_0, L_0) = \text{expit}(\epsilon C(A_0, L_0) + \hat{\alpha}_0 + \hat{\alpha}_1 A_0 + \hat{\alpha}_2 L_0). \tag{4}$$

The clever covariate is chosen so that the score (derivative of the log-likelihood) of the above logistic regression model at $\epsilon = 0$ equals the efficient influence function for the parameter (1) we are estimating. This is a "least-favorable" model, that is, a model that allows improvement in the direction in which the parameter we are estimating is most sensitive. 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, 2010).

The key step in targeted maximum likelihood estimation is updating a density estimate, such as the initial estimate described by the above logistic regression fits. A parametric model, with parameter ϵ, is constructed that (i) equals the current density estimate at $\epsilon = 0$, and (ii) has score at $\epsilon = 0$ equal to the efficient influence function. This parametric model is fit by maximum likelihood estimation, to obtain an updated density estimate. This process is repeated until convergence, that is, until the resulting estimate of ϵ is sufficiently close to 0. At that point, by properties (i) and (ii), the substitution estimator of the parameter at the current density estimate must (approximately) solve the efficient influence function estimating equation, with nuisance parameter evaluated at the current density estimate; this then can be used to prove double robustness and local efficiency of the estimator.

Here, we define the clever covariate to be $C(A_0, L_0) := A_0/\hat{p}(A_0 = 1|L_0)$, as derived in (Moore and van der Laan, 2007). We now update our estimate of $p(Y = 1|A_0, L_0)$, by fitting the logistic regression:

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

where the $\hat{\alpha}$ are considered fixed (they were computed above in (2)) 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

$$\text{expit}(\hat{\epsilon}C(1, L_0) + \hat{\alpha}_0 + \hat{\alpha}_1 + \hat{\alpha}_2 L_0). \tag{6}$$

Our estimate for the parameter (1) is the average over the distribution of L_0 in the data of our final estimate for $p(Y = 1|A_0 = 1, L_0)$ given in (6). Our estimate for this causal effect 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. This estimator is doubly robust, locally efficient, which means that if at least one of the models (2) or (3) is correctly specified, then the above estimator is consistent and asymptotically normal; if both models are correctly specified it is also efficient.

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, 2010). 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)

# 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:
```

```
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 causal effect of setting the treatment at both time points. It is straightforward to generalize the below discussion 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_i are the treatments, e.g. type of antiretroviral regimen. L_i are measurements such as CD4 count, viral load, etc. We let Y be the final outcome (death or not). So the variables we measure on each subject are: L_0, A_0, L_1, A_1, Y, where L_0 are baseline variables; A_0 is regimen just after resistance testing; 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 effect of setting treatment $A_0 = a_0$ and $A_1 = a_1$ on the mean of Y. 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 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. Under certain assumptions, the causal effect of setting $A_0 = a_0$ and $A_1 = a_1$ on the mean of Y is equal to the g-computation formula of Robins:

$$\sum_{l_0}\sum_{l_1} p(Y = 1 \mid A_1 = a_1, L_1 = l_1, A_0 = a_0, L_0 = l_0) \times \\ p(L_1 = l_1 \mid A_0 = a_0, L_0 = l_0) p(L_0 = l_0), \qquad (7)$$

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.

In what follows, we estimate the value of the above display at $a_0 = a_1 = 1$, that is, the causal effect of setting treatments A_0, A_1 to 1. Estimating the causal effect at other values of a_0, a_1 is similar. Targeted maximum likelihood estimation finds model fits for each of the conditional probabilities in the above formula, in a way targeted at estimating the overall parameter (7).

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) $p(L_0, A_0, L_1, A_1, Y)$.

Here is one example of a targeted maximum likelihood estimator for our problem. First, fit the following logistic regression models:

$$
\begin{aligned}
p(Y = 1 \mid A_1, L_1, A_0, L_0) &= \text{logit}^{-1}(\beta_0 + \beta_1 L_0 + \beta_2 A_0 + \beta_3 L_1 + \beta_4 A_1), && (8)\\
p(A_1 = 1 | L_1, A_0, L_0) &= \text{logit}^{-1}(\alpha_0 + \alpha_1 L_0 + \alpha_2 A_0 + \alpha_3 L_1), && (9)\\
p(L_1 = 1 \mid A_0, L_0) &= \text{logit}^{-1}(\gamma_0 + \gamma_1 L_0 + \gamma_2 A_0 + \gamma_3 L_0 A_0), && (10)\\
p(A_0 = 1 \mid L_0) &= \text{logit}^{-1}(\tau_0 + \tau_1 L_0), && (11)
\end{aligned}
$$

We denote the model fits by $\hat{p}$, e.g. $\hat{p}(A_0 = 1 \mid L_0 = l_0) = \text{logit}^{-1}(\hat{\tau}_0 + \hat{\tau}_1 l_0)$. Next, define the following "clever covariate," 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):

$$C_1(l'_0, a'_0, l'_1, a'_1) := \frac{1[a'_1 = 1]1[a'_0 = 1]}{\hat{p}(A_1 = 1 | L_1 = l'_1, A_0 = 1, L_0 = l'_0)\hat{p}(A_0 = 1 \mid L_0 = l'_0)}.$$

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 covariate C_1, using the previous 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

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

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

$$
\begin{aligned}
&\hat{p}_{\hat{\epsilon}_1}(Y = 1 \mid A_1, L_1, A_0, L_0) := \\
&\quad \text{logit}^{-1}(\hat{\epsilon}_1 C_1(L_0, A_0, L_1, A_1) + \hat{\beta}_0 + \hat{\beta}_1 L_0 + \hat{\beta}_2 A_0 + \hat{\beta}_3 L_1 + \hat{\beta}_4 A_1). \qquad (12)
\end{aligned}
$$

Next, define another clever covariate,

$$
\begin{aligned}
C_2(l'_0, a'_0) \;:=\; & \frac{1[a'_0 = 1]}{\hat{p}(A_0 = 1 \mid L_0 = l'_0)} \times \\
& \{\hat{p}_{\hat{\epsilon}}(Y = 1 \mid A_1 = 1, L_1 = 1, A_0 = 1, L_0 = l'_0) \\
& -\hat{p}_{\hat{\epsilon}}(Y = 1 \mid A_1 = 1, L_1 = 0, A_0 = 1, L_0 = l'_0)\}.
\end{aligned}
$$

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 previous 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

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

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

$$\hat{p}_{\hat{\epsilon}_2}(L_1 = 1 \mid A_0, L_0) := \text{logit}^{-1}(\hat{\epsilon}_2 C_2(L_0, A_0) + \hat{\gamma}_0 + \hat{\gamma}_1 L_0 + \hat{\gamma}_2 A_0 + \hat{\gamma}_3 L_0 A_0). \qquad (13)$$

Lastly, we compute the substitution estimator for (7) at the above model fits. That is, we evaluate (7) at $a_0 = a_1 = 1$ by substituting estimated densities (12) and (13) for the true densities, and using the empirical distribution for L_0 (which assigns mass $1/n$ to each observation). That is, our final estimate of the mean of Y setting A_0, A_1 both equal to 1, is

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{l_1\in\{0,1\}} \hat{p}_{\hat{\epsilon}_1}(Y = 1 \mid A_1 = 1, L_1 = l_1, A_0 = 1, L_0 = L_0^{(i)}) \times$$
$$\hat{p}_{\hat{\epsilon}_2}(L_1 = l_1 \mid A_0 = 1, L_0 = L_0^{(i)}).$$

This estimator is doubly robust, locally efficient; this means that if the models (8) and (10) are correctly specified, or if the models (9) and (11) are correctly specified, then the above estimator is consistent and asymptotically normal; if all four models are correctly specified it is also efficient.

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 (2010). Simple, efficient estimators of treatment effects in randomized trials using generalized linear models to leverage baseline variables. *The International Journal of Biostatistics. (Submitted).*

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).