Statistical Methods for Causal Inference in Observational and Randomized Studies

Mark J. van der Laan[1], Maya L. Petersen[1], Sherri Rose[2]

[1] University of California, Berkeley School of Public Health
[2] Johns Hopkins Bloomberg School of Public Health

laan@berkeley.edu · mayaliv@berkeley.edu · srose@jhsph.edu
stat.berkeley.edu/∼laan/
works.bepress.com/maya_petersen/
drsherrirose.com

targetedlearningbook.com

September 27, 2011

DAY TWO: LECTURE TWO

Other Target Parameters

Effect Modification

In many applications, one would like to estimate treatment specific mean of an outcome conditional on a user supplied baseline covariate (e.g. genetic profile).

In particular, this yields the treatment effect as a function of the baseline covariate.

Effect Modification: Examples

For example, one may be interested in these questions:

- What is the effect of an antidepressant medication on Hamilton Depression Rating Scale (HAM-D) score for those who enter a study with severe depression, and for those who enter with moderate depression?
- What is the effect of a cancer therapy for those who test positive for over-expression of a particular gene and for those who test negative for overexpression of that gene?
- What is the impact of low adherence to antiretroviral therapy on viral load for HIV-positive individuals who have just achieved viral suppression and for those who have maintained continuous viral suppression for 1 year?

High-Dimensional Exposure/Treatment

One might be interested in the effect of a continuous treatment such as the dose of a drug.

In addition, one might be interested in the effect of a multiple component treatment: e.g., drug1 and drug2, or drug and dose of drug.

Dynamic treatments

One may be interested in the effect of a rule for assigning a drug or dose of drug in response to characteristics of the subject.

For example, the class of dynamic treatments might be defined as: Treat if the patient's CD4-count drops below θ.

Stochastic Interventions

One may want to know the effect of a class of stochastic interventions defined as "assign a uniformly distributed dose between two values if the biomarker exceeds value θ".

Statistical Model

We assume a nonparametric statistical model for P_0; that is, we put no restrictions on the true data-generating distribution. The likelihood of the data at a candidate probability distribution P can be written

$$\prod_{i=1}^{n} P(Y_i, A_i, V_i, W_i) = \prod_{i=1}^{n} P_Y(Y_i \mid A_i, V_i, W_i) P_A(A_i \mid V_i, W_i) P_{V,W}(V_i, W_i).$$

Causal parameter of Interest

Let Y_a denote the potential outcome that would have been observed had treatment been at level $a \in \mathcal{A}$. We'd like to learn the probability that $Y_a = 1$, within strata $V = v$, that is

$$P(Y_a = 1 \mid V = v), a \in \mathcal{A}, v \in \mathcal{V}.$$

Identifiability: Statistical Parameter of Interest

We also would like to express the above display as a mapping from the distribution of the observed data. We make the following assumptions, which we use to connect the potential outcomes to the observed data:

- Time-ordering assumption: W, V precede A, which precedes Y;
- Consistency assumption: For all $a \in \mathcal{A}$, $Y = Y_a$ on the event $A = a$;
- Randomization assumption (no unmeasured confounders): $\{Y_a\}_{a \in \mathcal{A}} \perp\!\!\!\perp A \mid W, V$; and
- Positivity assumption: $P(A = a \mid W = w, V = v) > 0$ for all $a \in \mathcal{A}$ and all (w, v) in the support of P_0.

Statistical Parameter of Interest

Under these assumptions, we can equate function

$$P(Y_a = 1 \mid V = v), a \in \mathcal{A}, v \in \mathcal{V}.$$

of the potential outcomes we are interested in with a mapping from the distribution of the observed data, as follows:

$$P(Y_a = 1 \mid V = v) = E_{W|V=v} P(Y = 1 \mid A = a, V = v, W), a \in \mathcal{A}, v \in \mathcal{V},$$

where $E_{W|V=v}$ is expectation with respect to the distribution of baseline variables W given $V = v$.

Statistical Parameter of Interest

We define our parameter of interest $\Psi(P)$ to be the mapping from the observed data distribution given on the right-hand side of the previous display:

$$\Psi(P)(a, v) = E_{W|V=v} P(Y = 1 \mid A = a, V = v, W).$$

If A and V each had only a couple levels, we could estimate $\Psi(P_0)(a, v)$ (where P_0 is the true, unknown data-generating distribution) directly for each value of a and v.

The TMLE: Initial Estimator

We could fit a logistic regression model for $P_0(Y \mid A, V, W)$ such as

$$P(Y = 1 \mid A, V, W) = \text{expit}\left(\alpha_0 + \alpha_1 A + \alpha_2 V + \alpha_3 W\right).$$

One can also use super learning to obtain a data adaptive estimator of $P_0(Y = 1 \mid A, V, W)$.

Denote the fit by $\bar{Q}_n(Y = 1 \mid A, V, W)$. Our initial estimator Q_n^0 is defined as the pair $\left(\bar{Q}_n(Y = 1 \mid A, V, W), Q_{V,W,n}(V, W)\right)$.

The TMLE: Efficient Influence Curve

The efficient influence curve is (up to a normalizing constant)

$$\begin{aligned} D_{0,0}(Y,A,V,W) &= I(A=a,V=v)\left(\frac{Y-P(Y=1\mid A=a,V=v,W)}{P(A=a\mid V=v,W)}\right) \\ &\quad +I(V=v)[P(Y=1\mid A=a,V=v,W)-\Psi(P)(a,v)]. \end{aligned}$$

For practical identifiability, one wants a nicely bounded efficient influence curve as a function of O: Thus one needs that $P(A=a\mid V=v,W)$ is bounded away from zero.

The TMLE: Least Favorable Submodel

We now construct a parametric model $\{P_n^0(\epsilon) : \epsilon\}$ that 1) contains the initial estimator $P_n^0 = (Q_n^0, g_n)$ at $\epsilon = 0$ and (2) has a score at $\epsilon = 0$ whose linear span contains the efficient influence curve at $P_n^0 = (Q_n^0, g_n)$. To do this, we first define the clever covariate $H_1^*(A, V, W)$ for fluctuation of the outcome-regression, and function $H_2^*(V, W)$ for fluctuation of the distribution of (V, W):

$$H_1^*(A, V, W) = \frac{I(A = 0, V = 0)}{g_n(A = 0 \mid V = 0, W)}$$

and

$$H_2^*(V, W) = I(V = 0)[\bar{Q}_n(Y = 1 \mid A = 0, V = 0, W) - \Psi(Q_n^0)(0, 0)].$$

The TMLE

Let $\epsilon = (\epsilon_1, \epsilon_2)$. Define the parametric model $\{P(\epsilon) : \epsilon\}$:

$$\begin{aligned} P_n^0(\epsilon)(Y = 1 \mid A, V, W) &= \text{expit}\left(\epsilon_1 H_1^*(A, V, W) + \text{logit}\left(\bar{Q}_n(Y = 1 \mid A, V, W)\right)\right) \\ P_n^0(\epsilon)(A \mid V, W) &= g_n(A \mid V, W), \\ P_n^0(\epsilon)(V, W) &= s_{\epsilon_2} \exp(\epsilon_2 H_2^*(V, W)) Q_{V,W,n}(V, W), \end{aligned}$$

where the constant $s_{\epsilon_2} = 1/[\frac{1}{n}\sum_{i=1}^n \exp(\epsilon_2 H_2^*(V_i, W_i))]$ is chosen such that $P(\epsilon)(V, W)$ integrates to 1 for each ϵ.

The TMLE: TMLE-update step

We fit the above parametric model using maximum likelihood estimation to get estimates $\epsilon_n = (\epsilon_{1,n}, \epsilon_{2,n})$ of (ϵ_1, ϵ_2). Since the empirical distribution of (V, W) is an NPMLE, we have $\epsilon_{2,n} = 0$.

To obtain $\epsilon_{1,n}$, fit the logistic regression model on the previous slide, which has a single term (H_1^*) and offset equal to logit $\left(\bar{Q}_n(Y = 1 \mid A, V, W)\right)$.

Our final estimator for the relevant part Q_0 of the density of the data-generating distribution is

$$Q_n^* = \left(P_n^0(\epsilon_{1,n})(Y = 1 \mid A, V, W), Q_{V,W,n}\right).$$

The TMLE: Plug-in

Lastly, we compute the substitution estimator $\Psi(Q_n^*)(a, v)$:

$$\psi_n(a, v) = \frac{1}{\sum_{i=1}^n I(V_i = v)} \sum_{i=1}^n I(V_i = v) \bar{Q}_n^*(Y = 1 \mid A = a, V = v, W_i).$$

Unstable Estimator/ Practical Violation of Positivity Assumption

For a given treatment level a and covariate value v, the TMLE above for the parameter $\psi_0(a, v)$ defined on the previous slide involves the clever covariate:

$$\frac{I(A = a, V = v)}{g_n(a \mid v, W)}.$$

This estimator may become unstable if there are few subjects in the sample with $A = a$ and $V = v$.

The variance of the estimator will depend on the number of subjects in the category defined by $A = a$ and $V = v$.

One can consider two possible approaches for dealing with this.

Marginal Structural Model

The first approach is to assume a model m_β for the parameter $\psi_0(a, v)$ such as:

$$\text{logit } \psi_0(a, v) = \beta_0(a, v).$$

Such a model allows one to focus on estimating the parameter β_0, and the TMLE of β_0 will smooth across all the observations.

However, this requires making a model assumption (also restricting the statistical model!), and if this model assumption is incorrect (i.e., if there is model misspecification, which may be difficult to rule out), then β_0 (and thereby ψ_0) is not defined.

Defining Summary Measure: Working Marginal Structural Model

The second approach is to define our target parameter as a summary measure of the parameters $\{\psi_0(a, v) : a, v\}$.

For example, for a given treatment a, one could define our target parameter as the minimizer (β_0, β_1) of the expectation (with respect to the true data-generating distribution) of the squared residuals $(\psi_0(a, V) - \beta_0 - \beta_1 V)^2$.

In this case $\beta_0 + \beta_1 V$ represents the least squares projection of the true treatment-specific mean at level a as a function of V onto a linear trend.

Working Marginal Structural Model

The choice of working marginal structural model, such as the linear model $\beta_0 + \beta_1 V$, defines the target parameter of interest, but it does not represent a causal or statistical assumption.

Working Marginal Structural Model

The parameter $\Psi(P)$ is now well defined for any probability distribution P.

One could also define a whole collection of such summary measures as target parameters, thereby allowing the investigation of a whole collection of features of the true response curve $\psi_0(a, v)$ as a function of a and v.

Marginal Structural Working Model: Statistical Target Parameter

The parameter we will estimate is

$$\psi_0 = \arg\max_{\Psi'} \sum_{a \in \mathcal{A}} E_{P_0} h(a, V) \log \left[m(a, V, \Psi')^{Y_a} (1 - m(a, V, \Psi'))^{1 - Y_a} \right],$$

for some bounded, measurable weight function $h(a, V) \geq 0$ that we specify.

When the model m is correctly specified, this parameter yields $E(Y_a \mid V)$, and when m is misspecified, it represents the weighted-log-likelihood projection of this true dose-response curve onto the working model.

Statistical Target Parameter

In the above definition ψ_0, one can replace Y_a by $P_0(Y = 1 \mid A = a, V, W)$, and this ψ_0 is also the unique solution to

$$\sum_{a \in \mathcal{A}} E_{P_0} h(a, V)(P_0(Y = 1 \mid A = a, V, W) - m(a, V, \Psi'))(1, a_1, a_2, a_3, V)' = 0.$$

This defines a mapping from the distribution of the observed data.

The TMLE: Initial Estimator

We fit a logistic regression model to obtain an estimator for the first component $\bar{Q}_0$ of Q_0 and use the empirical distribution as estimator for the second component of Q_0. The resulting initial estimator Q_n^0 is denoted by $\big(\bar{Q}_n(Y = 1 \mid A, V, W), Q_{V,W,n}(V, W)\big)$.

We fit a multinomial logistic regression model for $P_0(A \mid V, W)$.

The TMLE: Efficient Influence Curve

To compute the optimal fluctuation submodel for the TMLE, we need the efficient influence curve for the parameter Ψ in the nonparametric model. The efficient influence curve is (up to a normalizing matrix) given by

$$D^*(P)(Y,A,V,W) = \left[\frac{h(A,V)(Y - P(Y=1 \mid A,V,W))}{P(A \mid V,W)} (1,A_1,A_2,A_3,V)' \right.$$
$$\left. + \sum_{a \in \mathcal{A}} h(a,V) \left(P(Y=1 \mid A=a,V,W) - m(a,V,\Psi') \right) (1,a_1,a_2,a_3,V)' \right].$$

For practical identifiabiliy, one wants $\max_{a\in\mathcal{A}} h(a,V)/P(A=a \mid V,W)$ to be nicely bounded. One may select h so that this is a reasonable assumption.

The TMLE: Submodel for fluctuation

We now construct a parametric model $\{P_n^0(\epsilon) : \epsilon\}$ that (1) contains the initial estimator (Q_n^0, g_n) at $\epsilon = 0$ and (2) has a score at $\epsilon = 0$ whose linear span contains the efficient influence function at (Q_n^0, g_n). To do this, we first define the clever covariates $H_1^*(A, V, W)$ and $H_2^*(V, W)$:

$$H_1^*(A, V, W) = \frac{h(A, V)}{g_n(A \mid V, W)}(1, A_1, A_2, A_3, V)'$$

and

$$\begin{aligned} H_2^*(V, W) = \sum_{a \in \mathcal{A}} h(a, V)(\bar{Q}_n(Y = 1 \mid A = a, V, W) \\ -m(a, V, \Psi'(Q_n^0)))(1, a_1, a_2, a_3, V)'. \end{aligned}$$

Here H_1^* and H_2^* are vectors.

The TMLE: Updating Step

Let $\epsilon = (\epsilon_1, \epsilon_2)$, where ϵ_1 and ϵ_2 are each row vectors with five components (so as to have the same length as H_1^* and H_2^*, respectively).

Define the parametric model $\{P_n^0(\epsilon) : \epsilon\}$:

$$\begin{aligned}
P_n^0(\epsilon)(Y = 1 \mid A, V, W) &= \text{expit}\left(\epsilon_1 H_1^*(A, V, W) + \text{logit}\left(\bar{Q}_n(Y = 1 \mid A, V, W)\right)\right) \\
P_n^0(\epsilon)(A \mid V, W) &= g_n(A \mid V, W), \\
P_n^0(\epsilon)(V, W) &= s_{\epsilon_2} \exp(\epsilon_2 H_2^*(V, W)) Q_{V,W,n}(V, W),
\end{aligned}$$

where the constant $s_{\epsilon_2} = 1/[\frac{1}{n}\sum_{i=1}^{n} \exp(\epsilon_2 H_2^*(V_i, W_i))]$ is chosen such that $P_n^0(\epsilon)(V, W)$ integrates to 1 for each ϵ.

The TMLE: Updating Step

We fit the above parametric model using maximum likelihood estimation to get the estimate $\epsilon_n = (\epsilon_{1,n}, \epsilon_{2,n})$ of (ϵ_1, ϵ_2). We have $\epsilon_{2n} = 0$ as before. $\epsilon_{1,n}$ can be obtained by fitting the logistic regression model, which has one term for each component of H_1^*, and offset equal to $\text{logit}\left(\bar{Q}_n(Y = 1 \mid A, V, W)\right)$.

Our final estimator for the relevant part Q_0 of the density of the observed data is

$$Q_n^* = P(\epsilon_n) = (P(\epsilon_{1,n})(Y = 1 \mid A, V, W), Q_{V,W,n})\,.$$

The TMLE: Plug-in

We compute the substitution estimator $\Psi'(Q_n^*)$, which solves

$$\sum_{a\in\mathcal{A}}\sum_{i=1}^{n} h(a,V_i)(\bar{Q}_n^*(Y=1\mid A=a,V_i,W_i) - m(a,V_i,\Psi'(Q_n^*))(1,a_1,a_2,a_3,V_i)' = 0.$$

The solution $\Psi'(Q_n^*)$ to the above equation can be computed using iteratively reweighted least squares, where the set of outcomes is $\bar{Q}_n^*(Y=1\mid A=a,V_i,W_i)$ for each $a\in\mathcal{A}$ and each subject i, which are regressed on the working model $m(a,V_i,\Psi')$ using weights $h(a,V_i)/[m(a,V_i,\Psi')(1-m(a,V_i,\Psi'))]$.

Practical Implementation of TMLE

This iteratively reweighted least squares solution can be implemented in the statistical programming language R with the generalized linear statistical model (glm) function. This involves first constructing a new data set where there are multiple rows for each subject, one for each possible level of treatment $a \in \mathcal{A}$.

Practical Implementation of TMLE

For subject i and treatment level $a \in \mathcal{A}$, the following entries make up the corresponding row of this new data set:

1. $\bar{Q}_n^*(Y = 1 \mid A = a, V_i, W_i)$ (which is the "outcome" in the new data set);
2. a (the adherence level under consideration; note that this is not the subject's observed adherence level);
3. V_i (the number of continuous months of past viral suppression);
4. $h(a, V_i)$ (the weight).

Practical Implementation of TMLE

One regresses the first column (the new "outcome") on the model $m(a, V_i, \Psi')$ using the glm function with family binomial and logistic link function and using weights $h(a, V_i)$ (from the fourth column of the new data set).

Even though the new "outcome" is not binary valued but lies in the interval $[0, 1]$, the glm function computes the desired iteratively reweighted least squares solution, as long as the algorithm converges.

Summary of Implementation of TMLE

We now summarize the steps in constructing the TMLE for the parameter.

1. Obtain the initial estimators of the conditional densities $P_0(Y = 1 \mid A, V, W)$ and $P_0(A \mid V, W)$.
2. Fit a logistic regression model for Y, with terms H_1^* and offset both depending on the initial density estimators and the formula for the efficient influence function for the parameter.
3. Use iterated reweighted least squares to solve

$$\sum_{a \in \mathcal{A}} \sum_{i=1}^{n} h(a, V_i)(\bar{Q}_n^*(Y = 1 \mid A = a, V_i, W_i) - m(a, V_i, \Psi(Q_n^*))(1, a_1, a_2, a_3, V_i)' =$$

yielding the final estimate ψ_n^*.

Dynamic Treatments

Let $W \rightarrow d(W)$ be a dynamic treatment rule so that $d(W) \in \mathcal{A}$. Let $\mathcal{D}$ be a collection of such dynamic treatments. One may be interested in

$$(E_P(Y_d \mid V = v) : d \in \mathcal{D}, v).$$

Positivity Assumption for Realistic Treatment Rules

The positivity assumption for identifiability of this cause curve is that $P(A = d(W) \mid W, V) > 0$ a.e.

Rules d can be selected so that this positivity assumption holds and such realistic rules might actually represent the true quantity of interest.

For example, for a given treatment level a one may define the rule d so that $d(W) = a$ if $P(A = a|W) > \delta > 0$ and $d(W) = a'$ otherwise, where a' is the level closest to a so that $P(A = a' \mid W) > \delta > 0$.

Working MSM for Dynamic Treatments

Given a working model m_β, one may define

$$\psi_0 = \arg\max_{\psi} \sum_{d \in \mathcal{D}} E_0 h(d, V) \log\{m_\beta(d, V)^{Y_d}(1 - m_\beta(d, V))^{1-Y_d}\},$$

where Y_d can be replaced by $E_0(Y \mid A = d(W), W, V)$. This defines the statistical target parameter on a nonparametric model.

The TMLE

The TMLE is defined as above for working MSM for static treatments, but now the clever covariate for updating an initial estimator $\bar{Q}_n^0$ of $E_0(Y \mid A, W, V)$ is given by:

$$H_1^*(A, V, W) = \sum_{d \in \mathcal{D}} \frac{h(d, V)}{P(A = d(W) \mid W, V)} \frac{\frac{d}{d\beta} m_\beta(d, V)}{m_\beta(1 - m_\beta)(d, V)}.$$

Concluding Remarks

- Working marginal structural models provide interesting summary measures of causal effects of static, dynamic, and stochastic interventions on an outcome of interest.
- These summary measures allow smoothing across treatment levels, and can be estimated with TMLE using standard regression or machine learning methodology.
- TMLE incorporates both an estimator of the outcome regression as well as an estimate of the propensity score/treatment mechanism.
- The TMLE are double robust and asymptotically efficient if both are consistently estimated.
- The above presented TMLE can also be applied to continuous outcomes $Y \in [0, 1]$, and thereby naturally handles bounded outcomes as well. Such TMLE respect the global constraints on the outcome, and are therefore more robust to practical violations of the positivity assumption.