Lecture 11: TMLE

Outline

- Intro to TMLE:
  - Properties
  - Implementation: TMLE for ATE estimand
- Some background on TMLE
  - Estimating Equations
  - Estimating Equations and influence curves
  - Efficient influence curve
  - TMLE solves estimating equation corresponding to efficient IC
- A-IPW: DR efficient estimating equation-based estimator
- TMLE in practice…

References

- TLB. Chapters 4-6
- Kennedy, 2017: https://arxiv.org/abs/1709.06418v1

The Roadmap

1. Specify **Causal Model** representing real background knowledge
2. Specify **Causal Question**
3. Specify **Observed Data** and link to causal model
4. **Identify** : Knowledge + data sufficient?
5. Commit to an **estimand** as close to question as possible, and a **statistical model** representing real knowledge.
6. **Estimate**
7. **Interpret** Results

Estimate the Chosen Parameter of the Observed Data Distribution

- For illustration we are focusing primarily on a single statistical estimation problem
  - $O=(W,A,Y)\sim P_0$
    - Statistical model is non- or semi-parametric
  - $\Psi(P_0)=E_{W,0}(E_0(Y|A=1,W)-E_0(Y|A=1,W))$
    - If W satisfies backdoor criteria, equal to the ATE
- Focusing on three classes of estimator
  - Simple substitution (G-comp)
  - IPTW
  - Today: Double Robust- Specifically, AIPW and TMLE

Overview of Estimators

- Each class of estimator requires for its implementation an estimator of a distinct factor of the observed data distribution
- Distribution of the observed data:
- $P_0(O)=P_0(W,A,Y) = P_0(W)\ P_0(A|W)\ P_0(Y|A,W)$

Different Estimators require estimators of distinct factors of the observed data distribution

- $P_0(O)=P_0(W,A,Y)=P_0(W)P_0(A|W)\ P_0(Y|A,W)$
- Simple substitution estimators
  - Also referred to as "G-computation" estimators
  - Actually, rather than full $P_0(Y|A,W)$, only require estimators of $E_0(Y|A,W)$ (and $P_0(W)$)
- Consistency depends on consistent estimation of $E_0(Y|A,W)$
  - Super Learning can help here, but…

IPTW Estimators

- $P_0(O)=P_0(W,A,Y)=P_0(W)P_0(A|W)\,P_0(Y|A,W)$

Inverse probability weighted estimators

- Consistency of IPTW estimators depends on consistent estimation of $g_0\,(A|W)=P_0(A|W)$
  - Super learning can help here, but….

Coming next: Double Robust Estimators

- $P_0(O)=P_0(W,A,Y)=P_0(Y|A,W)\ P_0(A|W)\ P_0(W)$

  Double Robust estimators:
  - A-IPTW
  - TMLE

- These asymptotic properties typically translate into lower bias and variance in finite samples
- Can integrate machine learning and still maintain valid statistical inference
  - Meaningful CIs and p values

Coming next: Double Robust Estimators

- $P_0(O)=P_0(W,A,Y)=P_0(Y|A,W)\ P_0(A|W)\ P_0(W)$

  Double Robust estimators:
  - A-IPTW
  - TMLE

- Implementation requires estimators of both $E_0(Y|A,W)$ and $g_0(A|W)$
- Consistent if either $E_0(Y|A,W)$ or $g_0(A|W)$ are estimated consistently

Coming next: Double Robust Estimators

- $P_0(O)=P_0(W,A,Y)=P_0(Y|A,W)\ P_0(A|W)\ P_0(W)$

Double Robust estimators:

- A-IPTW
- TMLE

- If both $E_0(Y|A,W)$ and $g_0(A|W)$ are estimated consistently (at rates faster than $n^{-1/4}$) then these estimators are efficient
  - Lowest asymptotic variance of any reasonable estimator
  - In semiparametric (or non-parametric) statistical model that makes assumptions, if any, only on $P_0(A|W)$

Targeted Maximum Likelihood Estimation

- TMLE is a general methodology
- As with other estimators, we will focus on estimation of the "G comp estimand" corresponding under causal assumptions to the ATE:

$$\Psi(P_0)=E_{W,0}(E_0(Y|A=1,W)-E_0(Y|A=0,W))$$

General Overview: TMLE

1. Estimate the portion of P_0 that the target parameter is a function of (i.e., estimate Q_0)
   - $\Psi(P_0)=\Psi(Q_0)$

- What is Q_0 for the G-comp estimand?
- How could you estimate it?

General Overview: TMLE

2. Update initial estimator of Q_0 to obtain targeted fit of Q_0

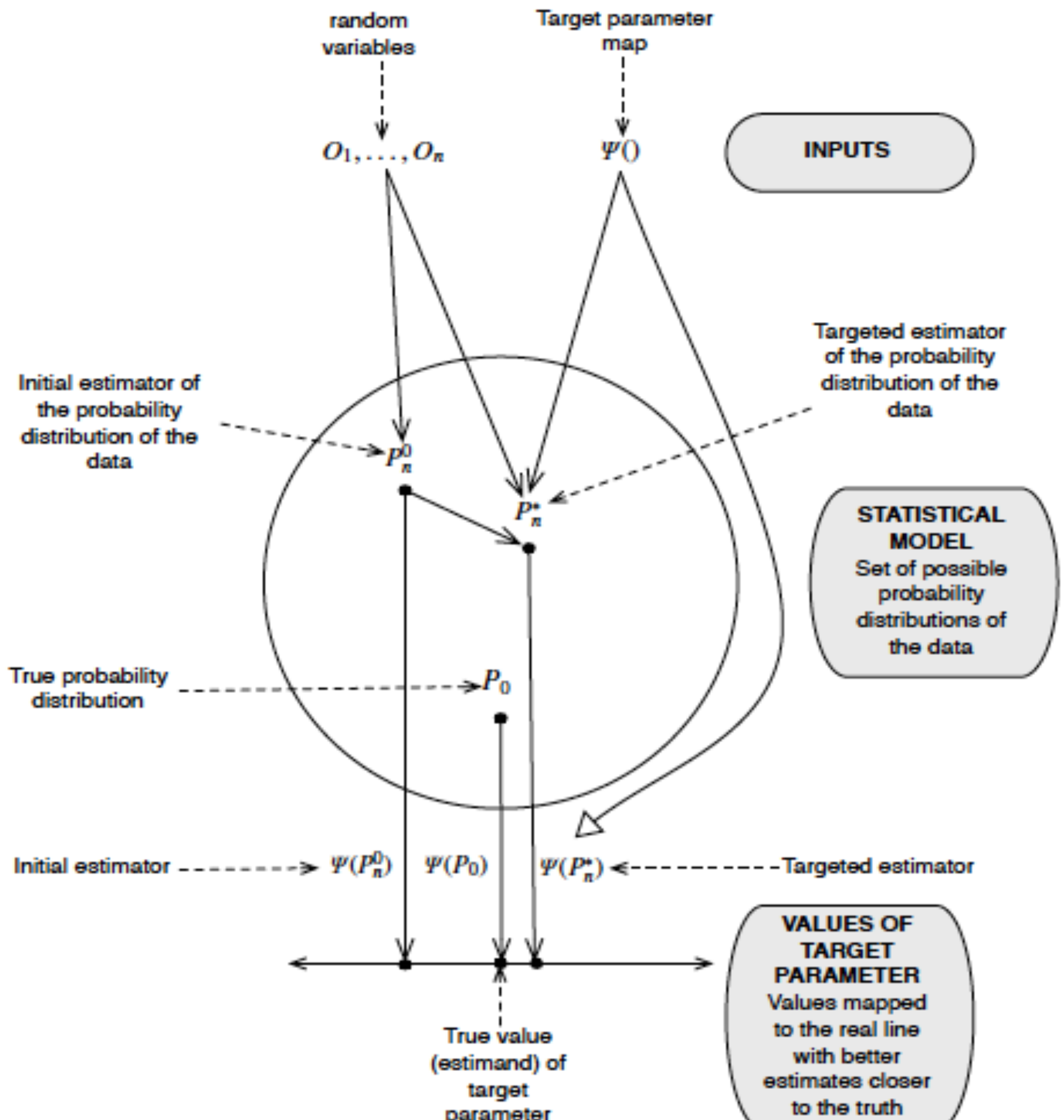
- Targeting makes use of information in P_0 beyond Q_0 to improve estimation of ψ_0
- Provides an opportunity to
  - Reduce asymptotic bias if initial initial estimator of Q_0 not consistent
  - Reduce finite sample bias
  - Reduce variance

General Overview: TMLE

3. Plug in updated (targeted) estimator of Q_0 into the parameter mapping Ψ to generate estimate

- What do we call this type of estimator?
- Assume you have the targeted fit of Q_0 (we haven't talked about how to get it yet).

  How would you estimate G comp estimand $\Psi(P_0)$?


random variables
Target parameter map
$O_1, \ldots, O_n$
$\Psi()$
INPUTS
Targeted estimator of the probability distribution of the data
Initial estimator of the probability distribution of the data
P_n^0
P_n^*
STATISTICAL MODEL
Set of possible probability distributions of the data
True probability distribution
P_0
Initial estimator
$\Psi(P_n^0)$
$\Psi(P_0)$
$\Psi(P_n^*)$
Targeted estimator
VALUES OF TARGET PARAMETER
Values mapped to the real line with better estimates closer to the truth
True value (estimand) of target parameter

Overview of TMLE for ATE estimand

1. Estimate $E_0(Y|A,W)$
   - Use machine learning to respect statistical model
   - Gives "best" estimate of $E_0(Y|A,W)$
2. Modify this initial estimate of $E_0(Y|A,W)$
   - Target it to give better estimate of $\Psi(P_0)=E_W(E_0(Y|A=1,W)-E_0(Y|A=0,W))$
   - This targeting requires estimation of $g_0(A|W)$
3. Implement substitution estimator with new targeted estimate of $E_0(Y|A,W)$
   - For the TMLE, generally have that

$$\sqrt{n}\left(\hat{\Psi}(P_n)-\Psi(P_0)\right)\to N(0,\sigma^2)$$

Step by Step Overview: TMLE

1. Estimate $E_0(Y|A,W) \equiv \bar{Q}_0(A,W)$
   – Eg using super learner

– Notation for this initial estimate of $E_0(Y|A,W)$:

$$\bar{Q}_n^0(A,W)$$

"n" because it is an estimate of the true parameter value

"0" refers to initial (non-targeted) estimate

2. Generate predicted values for Y for each individual, given that individual's A_i,W_i
   – For participant i: $\bar{Q}_n^0(A_i,W_i)$

Step by Step Overview: TMLE

3. Estimate treatment mechanism
   - $g_0(A|W)$
   - Eg using Super Learner

Step by Step Overview: TMLE

4. Use this estimate to create a new "clever covariate" $H_n(A,W)$ for each individual
   - For subject i

$$H_n(A_i, W_i) \equiv \left(\frac{I(A_i = 1)}{g_n(A_i = 1|W_i)} - \frac{I(A_i = 0)}{g_n(A_i = 0|W_i)} \right)$$

   - We will use this clever covariate to update our initial estimate

Step by Step Overview: TMLE

5. Update the initial estimate of $E_0(Y|A,W)$

- Run a logistic regression of Y_i on $H_n(A_i,W_i)$ using $\text{logit}(\bar{Q}_n^0(A_i,W_i))$ (predicted value Y for each person) as offset (suppressing intercept term)

$$\text{logit}(E(Y \mid A_i,W_i) = \text{logit}(\bar{Q}_n^0(A_i,W_i)) + \varepsilon H_n(A_i,W_i)$$

- Let ε_n denote the resulting MLE estimate of the coefficient ε on $H_n(A,W)$
- Updated estimate:

$$\bar{Q}_n^*(A,W) = \text{expit}\left(\text{logit}(\bar{Q}_n^0) + \varepsilon_n H_n(A,W)\right)$$

Why? Very Informal Intuition

- Want to move our initial estimate $\overline{Q}_n^0(A,W)$ closer to the truth $\bar{Q}_0(A,W) \equiv E_0(Y|A,W)$
- Why? Because our initial estimate was aimed at achieving optimal bias/variance tradeoff for full regression function $E_0(Y|A,W)$
  - Wrong bias variance tradeoff for the target parameter
    - Target parameter is lower dimensional- a single number, not a prediction for every (A,W) combination
- How? Need to do this in a targeted way
  - We want to change the initial estimate by fitting it to the data where it matters most for target parameter

Very Informal Intuition

- Not all deviations between initial estimate $\bar{Q}_n^0(A,W)$ and truth are equally important
  - With confounding, certain covariate/treatment combinations are underrepresented
    - Relative to ideal situation (from a causal perspective) in which covariate distributions are balanced across treatment levels –think of IPTW
  - A large deviation between our initial estimator and the truth for an (a,w) level for which g(a|W=w) is small is more important to our (causally motivated) target parameter than its frequency in the observed data reflects

Very Informal Intuition

- Need to make this explicit when we update
  - "Tell MLE" to give individuals with small predicted probability of observed treatment ($g_n(A|W)$ small) more weight when updating initial fit
- How can we do this?
- One option "clever covariate"

$$H_n(A_i, W_i) \equiv \left(\frac{I(A_i = 1)}{g_n(A_i = 1|W_i)} - \frac{I(A_i = 0)}{g_n(A_i = 0|W_i)} \right)$$

  - if $g_n(A|W)$ is small, absolute covariate value is big, and thus a small change in epsilon has a bigger impact on the fit

Step by Step Overview: TMLE

6. Calculate predicted values for each individual under each treatment level of interest using the updated estimate $\overline{Q}_n^*(A,W)$

- For each individual, set a=1 and a=0 and generate predicted outcome with updated estimate

$$\overline{Q}_n^*(1,W_i) = \text{expit}(\text{logit}(\overline{Q}_n^0(1,W_i)) + \varepsilon_n H_n(1,W_i))$$

$$\overline{Q}_n^*(0,W_i) = \text{expit}(\text{logit}(\overline{Q}_n^0(0,W_i)) + \varepsilon_n H_n(0,W_i))$$

if $g_n(0|W_i)$ is small,
Then $H_n(0,W_i)$ is big,
and initial fit is updated more

Step by Step Overview: TMLE

7. Estimate $\Psi(P_0)$ as the empirical mean of the predicted values of Y for a=1 and a=0, based on the updated fit

$$\hat{\Psi}_{TMLE}(P_n) = \frac{1}{n}\sum_{i=1}^{n}\left[\bar{Q}_n^*(1, W_i) - \bar{Q}_n^*(0, W_i)\right]$$

A bit more about why this update works…

1. Estimating functions and estimating equations
2. Link to influence curves
3. The efficient influence curve

- TMLE is a substitution estimator that also solves the estimating equation corresponding to the efficient influence curve

Brief intro to estimating equations

- An Estimating Function $D(O|\ \psi)$ is a function of the observed data and the (unknown) parameter of interest
  - Observe n i.i.d. copies of O_i, i=1,…n; $O \sim P_0$
  - Parameter of interest $\Psi(P_0)=\psi$
  - Unbiased estimating function: $E_0[D(O|\ \psi)]=0$
- Estimating Equation: $$0 = \frac{1}{n}\sum_{i=1}^{n} D(O_i|\psi)$$
- Estimator: ψ_n defined as the solution satisfying $$\frac{1}{n}\sum_{i=1}^{n} D(O_i|\psi_n) = 0$$

Simple example: Population mean

- Observe n i.i.d. copies of $O_i=Y_i$; $O \sim P_0$
- Parameter of interest $\Psi(P_0)= \psi =E_0(Y)$
- Let $D(O \mid \psi)=Y- \psi$
  - Note: $E_0[D(O \mid \psi)]=E_0(Y)- \psi=0$
- Estimating Equation: $$0 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \psi)$$
- Estimator $$\psi_n = \frac{1}{n}\sum_{i=1}^{n}(Y_i)$$
  - Sample mean as estimator of population mean can be understood as root of an estimating equation

IPTW estimator defined as solution to an estimating equation

- Observe n i.i.d. copies of $O_i=(W_i\ A_i\ Y_i)$; $O \sim P_0$
- Parameter of interest: $$\Psi(P_0) = E_0\left(\frac{I(A=a)}{g_0(A|W)}Y\right)$$
- Estimating function: $$D_{IPTW}(O|g,\psi) = \frac{I(A=a)}{g(A|W)}Y - \psi$$
  - Note: if treatment mechanism *g* is not known, then it is a "nuisance parameter" which must be estimated
- Estimating Equation: $$0 = \frac{1}{n}\sum_{i=1}^{n}\frac{I(A_i=a)}{g_n(A_i|W_i)}Y_i - \psi$$

(Non-stabilized) IPTW estimator: $$\psi_n = \frac{1}{n}\sum_{I=1}^{n}\frac{I(A_i=a)}{g_n(A_i|W_i)}Y_i$$

Influence Curves and Estimating Functions

- Recall: An estimator is asymptotically linear with influence curve $IC(O_i)$ if it satisfies

$$\psi_n - \psi = \frac{1}{n}\sum_{i=1}^{n} IC(O_i) + o_{P_0}\left(\frac{1}{\sqrt{n}}\right)$$

$E_0(IC(O))=0$
$Var(IC(O))$ Finite

Converges to 0 in probability as $n\to\infty$, even when multiplied by $\sqrt{n}$

- Because $E_0(IC(O))=0$, if we know the IC of an estimator, then we can use it as an estimating function
  - Estimating equation will be unbiased, up to a second order term

Example: IC of IPTW

- Assume g_0 is known, and strong positivity
- IC of the IPTW estimator:

$$D_{IPTW}(O|g_0,\psi) = \frac{I(A=a)}{g_0(A|W)}Y - \psi$$

  – Note: $E_0 D_{IPTW}(O|g_0, \psi)=0$

$$\psi_n - \psi = \frac{1}{n}\sum_{i=1}^{n} D(O_i|g_0,\psi)$$

Exact equality-
no remainder term

$$= \frac{1}{n}\sum_{i=1}^{n} \frac{I(A_i=a)}{g_0(A_i|W_i)}Y_i - \psi$$

  – For known g_0, variance of the IPTW estimator well-approximated by $\mathrm{var}[D(O_i|g_0, \psi_n)]/n$

Example: IC of IPTW

- If g_0 is estimated using a correctly specified parametric model, the IC that treats g_0 as known gives conservative variance estimates (overestimates variance)
- Why?

  IC of IPTW estimator with g_0 estimated=

  IC of IPTW estimator with g_0 known – *projection*
- *Estimating g_0 (using a correctly specified parametric model) reduces the variance of the IC and thus of the IPTW estimator*

Influence Curves vs. The Efficient Influence Curve

- For a given a statistical estimation problem:
  - n i.i.d. copies of O_i, $O \sim P_0 \in \mathcal{M}$
  - Target parameter $\Psi(P_0)=\psi$

1. Influence curves (or influence functions) are *estimator- specific*
   - Each asymptotically linear estimator has an influence curve
   - Influence curve teaches us about the asymptotic variance of the estimator
2. The efficient influence curve is *parameter-specific*
   - Teaches us about the asymptotic variance of the **most efficient** regular asymptotically linear estimator for that parameter
   - i.e. the estimator with the lowest asymptotic variance

Efficient Influence Curve

- An estimator is efficient if and only if it is asymptotically linear with influence curve the efficient influence curve $D^*(P_0)$:

$$\hat{\Psi}(P_n) - \Psi(P_0) = \frac{1}{n}\sum_{i=1}^{n} D^*(P_0)(O_i) + o_P\left(1/\sqrt{n}\right)$$

  - Efficient influence curve needs to be derived for a given estimation problem
- An efficient estimator needs to solve the estimating equation corresponding to efficient influence curve (up to second order term)

$$0 = \frac{1}{n}\sum_{i=1}^{n} D^*(P)(O_i)$$

TMLE solves efficient IC equation

- TMLE solves $0 = \frac{1}{n}\sum_{i=1}^{n} D^*(P_n^*)(O_i)$
  - Efficient influence curve:

$$D^*(P) = \underbrace{\left[\frac{A}{g(A \mid W)} - \frac{1-A}{g(0 \mid W)}\right]\left[Y - \bar{Q}(A,W)\right]}_{\textbf{a}} + \underbrace{\bar{Q}(1,W) - \bar{Q}(0,W) - \psi}_{\textbf{b}}$$

  - Stage 2 targeting fits ε by maximum likelihood
    - MLE solves score equation $\sum_{i=1}^{n} H_n(A_i,W_i)\left[Y_i - \bar{Q}_n^*(A_i,W_i)\right] = 0$
    - We defined our parameter-specific H_n and fit with MLE to ensure that empirical mean of **a** equals 0
  - As a substitution estimator, $\psi_n^{TMLE} = \frac{1}{n}\sum_{i=1}^{n}\left[\bar{Q}_n^*(1,W_i) - \bar{Q}_n^*(0,W_i)\right]$ thus empirical mean of **b** equals 0

Influence curve-based Inference

- Under conditions (see Ch 27 TLB) TMLE is asymptotically linear estimator
- If g_0 and Q_0 are estimated consistently, then the influence curve of the resulting TMLE equals the Efficient Influence Curve

$$D^*(P_0)(O) = \left(\frac{I(A=1)}{g_0(1|W)} - \frac{I(A=0)}{g_0(0|W)}\right)\big(Y - \bar{Q}_0(A,W)\big) + \bar{Q}_0(1|W) - \bar{Q}_0(0,W) - \psi_0$$

- Depends on unknown nuisance parameters g_0 and Q_0
  - Can estimate the influence curve of the TMLE as:

$$IC_n(O) = \left(\frac{I(A=1)}{g_n(1|W)} - \frac{I(A=0)}{g_n(0|W)}\right)\big(Y - \bar{Q}^*_n(A,W)\big) + \bar{Q}^*_n(1|W) - \bar{Q}^*_n(0,W) - \psi_n$$

- If g_0 is estimated consistently (with MLE) but Q_0 is not then this provides conservative approximation of the IC
  - i.e. can use it to get conservative variance estimate

Influence curve-based Inference

- Variance of an asymptotically linear estimator is well-approximated by the variance of its Influence curve/n

1. (Conservatively) estimate the TMLE Influence Curve by plugging in estimates of g_n and Q_n
2. To estimate variance of the estimator: take sample variance of the estimated influence curve and divide by sample size

- 95% CI: $\psi_n(Q_n^*) \pm 1.96\hat{\sigma}/\sqrt{n}$

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} \hat{D}^{*2}(P_n^*)(O_i)$$

Augmented IPTW

- Efficient and double robust
  - Like TMLE: Solves the the estimating equation corresponding to the efficient influence curve
- Defined as a solution to an estimating equation
  - Unlike TMLE: Not a substitution estimator

– Define estimating function:

$$D^*(O|Q,g,\psi) = \left(\frac{I(A=1)}{g(1|W)} - \frac{I(A=0)}{g(0|W)}\right)\left(Y - \bar{Q}(A,W)\right) + \bar{Q}(1|W) - \bar{Q}(0,W) - \psi$$

– Estimate g and Q and solve estimating equation for ψ

$$0 = \frac{1}{n}\sum_{i=1}^{n}\left[\left(\frac{I(A_i=1)}{g_n(1|W_i)} - \frac{I(I=0)}{g_n(0|W_i)}\right)\left(Y_i - \bar{Q}_n(A_{i,i}\,)\right) + \bar{Q}_n(1|W_i) - \bar{Q}_n(0,W_i) - \psi\right]$$

$$\psi_n = \frac{1}{n}\sum_{i=1}^{n}\left[\left(\frac{I(A_i=1)}{g_n(1|W_i)} - \frac{I(I=0)}{g_n(0|W_i)}\right)\left(Y_i - \bar{Q}_n(A_{i,i}\,)\right) + \bar{Q}_n(1|W_i) - \bar{Q}_n(0,W_i)\right]$$

Why we might prefer TMLE to other double robust estimators

- As a substitution estimator, automatically respects the bounds of the model
  - This is important when there are near positivity violations, i.e. g_0 is close to zero
    - Can improve stability
  - Nera positivity violations can still impact the performance of TMLE…
- In general, estimating equations…
  - Might not have a solution
  - Might only have a solution outside parameter space
  - Might have multiple solutions, with no criterion to choose between them…

TMLE: Some take home messages

- TMLE is Double Robust: Consistent if either g_0 or Q_0 are estimated consistently
- TMLE is efficient if g_0 and Q_0 are both estimated consistently at a reasonable rate
- This can translate into real bias and variance improvement
  - Reduce asymptotic bias if initial initial estimator of Q_0 not consistent
  - Reduce finite sample bias
  - Reduce variance

TMLE: Some take home messages

- Use data-adaptive estimation (Super Learning) for g and Q
  - Asymptotic linearity relies on bias disappearing at a fast enough rate
  - Influence curve-based inference relies on g_0 being estimated consistently
    - Conservative variance estimate
  - Good estimation of both g_0 and Q_0 gives us efficiency

TMLE: Beyond simple single time point…

- TMLE is a general method; broad applications
  - Longitudinal problems with time-dependent confounding
  - Parameters of (longitudinal) marginal structural models
  - Dynamic regimes (personalized treatment/adaptive strategies)
  - Informative censoring
  - RCTs (including SMART designs) for improved efficiency
- Estimands, estimators and implementation differ
- R packages implementing all of the above are available (ltmle, tmle, SuperLearner)

Example: Alternative TMLE

- Can also target initial estimate $\overline{Q}_n^0$ by running an intercept-only weighted logistic regression with:
  - Outcome: Y
  - Offset: $\text{logit}(\overline{Q}_n^0)$
  - Weight: $H_n(A,W) = \dfrac{I(A=1)}{g_n(A \mid W)}$
- i.e. have moved the "clever covariate" to the weights
  - This has benefits in face of positivity violations
  - This is the option implemented in ltmle package

General TMLE Procedure

1. Identify the “hardest” parametric submodel to fluctuate initial estimate of P_0
   – Small fluctuation -> maximum change in target
2. Identify optimal magnitude of fluctuation by MLE
3. Apply optimal fluctuation to Initial estimate to obtain 1st-step TMLE
4. Repeat until incremental fluctuation is zero
   – 1-step convergence guaranteed in some important cases
5. Final probability distribution solves efficient influence curve equation
   – Basis for asymptotic linearity, normality, and efficiency.
   – Confers double robustness, or, more general, makes bias a second order term.