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

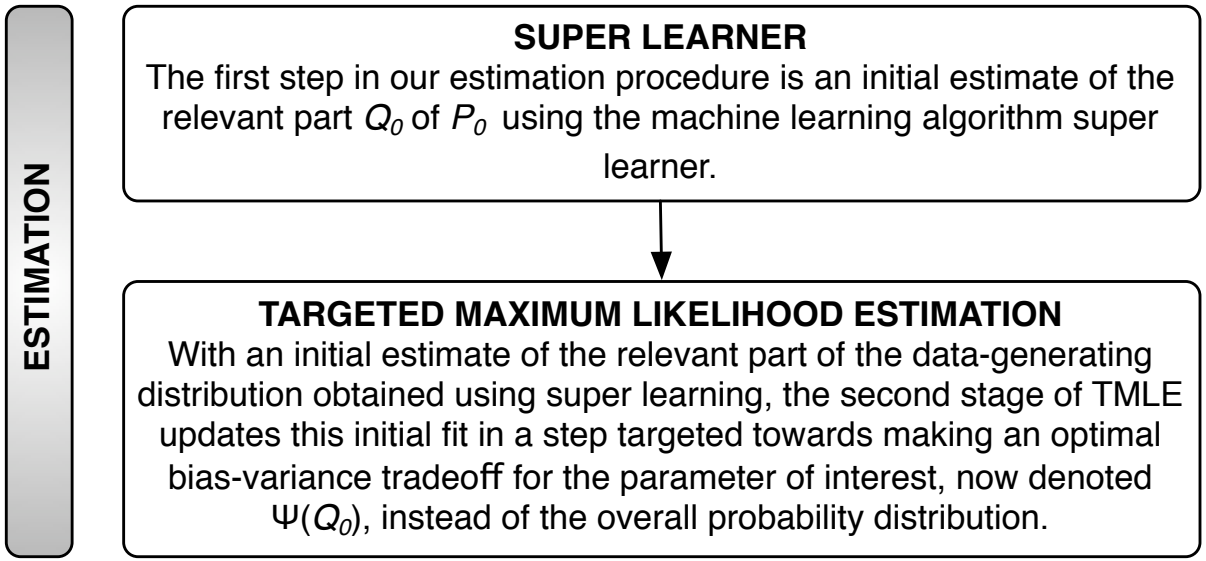
targetedlearningbook.com

September 26, 2011

DAY ONE: LECTURE THREE

Estimation: MLE, (A-)IPTW, TMLE

Road map - Estimation



Landscape

- **MLE**
- **Estimating-Equation-Based Methods**
- **TMLE**

MLE

A maximum likelihood estimator for a parametric statistical model $\{p_\theta : \theta\}$ is defined as a maximizer over all densities in the parametric statistical model of the empirical mean of the log density:

$$\theta_n = \arg\max_\theta \sum_{i=1}^{n} \log p_\theta(O_i).$$

MLE

The $L(p)(O) = -\log p(O)$ is called a loss function at candidate density p for the true density p_0 since its expectation is minimized across all densities p by the true density $p = p_0$.

This minimization property of the log-likelihood loss function is the principle behind maximum likelihood estimation providing the basis for establishing that maximum likelihood estimators for correctly specified statistical models approximate the true distribution P_0 for large sample size.

MLE

This discussion can be equally applied to the case where $L(p)(O)$ is replaced by any other loss function $L(Q)$ for a relevant part Q_0 of p_0, satisfying that $E_0 L(Q_0)(O) \leq E_0 L(Q)(O)$ for each possible Q. In that case, we might call this estimator a minimum-loss-based estimator.

TMLE incorporates this case as well, in which it could be called targeted minimum-loss-based estimation (still abbreviated as TMLE).

We focus our comparison on the log-likelihood loss function and will thereby refer to MLE

MLE

$\Psi(P_0)$ for the causal risk difference can be written as the g-formula:

$$\Psi(P_0) = \sum_w \left[\sum_y y P_0(Y = y \mid A = 1, W = w) - \sum_y y P_0(Y = y \mid A = 0, W = w) \right] P_0(W = w),$$

where

$$P_0(Y = y \mid A = a, W = w) = \frac{P_0(W = w, A = a, Y = y)}{\sum_y P_0(W = w, A = a, Y = y)}$$

is the conditional probability distribution of $Y = y$, given $A = a$, $W = w$, and

$$P_0(W = w) = \sum_{y,a} P_0(W = w, A = a, Y = y).$$

MLE

Maximum-likelihood-based substitution estimators of the g-formula are obtained by substitution of a maximum-likelihood-based estimator of Q_0 into the parameter mapping $\Psi(Q_0)$.

The marginal distribution of W can be estimated with the nonparametric maximum likelihood estimator, which happens to be the empirical distribution that puts mass $1/n$ on each W_i, $i = 1, \ldots, n$.

MLE

Maximum-likelihood-based substitution estimators will be of the type

$$\psi_n = \Psi(Q_n) = \frac{1}{n}\sum_{i=1}^{n}\{\bar{Q}_n(1, W_i) - \bar{Q}_n(0, W_i)\},$$

where this estimate is obtained by plugging in $Q_n = (\bar{Q}_n, Q_{W,n})$ into the mapping Ψ.

MLE

MLE using stratification. The simplest maximum likelihood estimator of $\bar{Q}_0$ stratifies by categories or possible values for (A, W). One then simply averages across the many categories.

In most data sets, there will be a large number of categories with few or zero observations. One might refer to this as the curse of dimensionality, making the MLE for nonparametric statistical models typically ill defined, and an overfit to the data resulting in poor finite sample performance. One can refer to this estimator as the nonparametric MLE (NPMLE).

MLE

MLE after dimension reduction: propensity score methods. To deal with the curse of dimensionality, one might propose a dimension reduction W^r of W and apply the simple MLE to the reduced-data structure (W^r, A, Y).

However, such a dimension reduction could easily result in a biased estimator of $\Psi(Q_0)$ by excluding confounders. One can show that a sufficient confounder is given by the propensity score $g_0(1 \mid W) = P_0(A = 1 \mid W)$, allowing one to reduce the dimension of W to only a single covariate, without inducing bias.

MLE

MLE after dimension reduction: propensity score methods. A maximum likelihood estimator of $E_0(Y \mid A, W^r)$ can then be applied, where $W^r = g_0(1 \mid W)$, using stratification.

For example, one creates five categories for the propensity score, thereby creating a total of ten categories for (A, W^r), and estimates $E_0(Y \mid A, W^r)$ with the empirical average of the outcomes within each category. Of course, this propensity score is typically unknown and will thus first need to be estimated from the data.

MLE

MLE using regression in a parametric working model. $\bar{Q}_0(A, W)$ is estimated using regression in a parametric working (statistical) model and plugged into the formula given previously.

MLE

ML-based super learning. We estimate $\bar{Q}_0$ with the super learner, in which the collection of estimators may include stratified maximum likelihood estimators, maximum likelihood estimators based on dimension reductions implied by the propensity score, and maximum likelihood estimators based on parametric working models, beyond many other machine learning algorithms for estimation of $\bar{Q}_0$.

We will discuss super learning on **DAY TWO.**

Estimating Equation Methods

Estimating-equation-based methodology for estimation of our target parameter $\Psi(P_0)$ includes inverse probability of treatment-weighted (IPTW) estimators and augmented IPTW (A-IPTW) estimators. These methods aim to solve an estimating equation in candidate ψ-values.

Estimating Equation Methods

An estimating function is a function of the data O and the parameter of interest. If $D(\psi)(O)$ is an estimating function, then we can define a corresponding estimating equation:

$$0 = \sum_{i=1}^{n} D(\psi)(O_i),$$

and solution ψ_n satisfying $\sum_{i=1}^{n} D(\psi_n)(O_i) = 0$.

Estimating Equation Methods

Most estimating functions for ψ will also depend on an unknown "nuisance" parameter of P_0. So we might define the estimating function as $D(\psi, \eta)$, where η is a candidate for the nuisance parameter. Given an estimator η_n of the required true nuisance parameter η_0 of P_0, we would define the estimating equation as

$$0 = \sum_{i=1}^{n} D(\psi, \eta_n)(O_i),$$

with solution ψ_n satisfying $\sum_{i=1}^{n} D(\psi_n, \eta_n)(O_i) = 0$.

Estimating Equation Methods

When the notation $D^*(\psi_0, \eta_0)$ is used for the estimating function $D(\psi_0, \eta_0)$, $D^*(\psi_0, \eta_0)$ is an estimating function implied by the efficient influence curve.

An efficient influence curve is $D^*(P_0)(O)$, i.e., a function of O, but determined by P_0, and may be abbreviated $D^*(P_0)$ or $D^*(O)$. An optimal estimating function is one such that $D(\psi_0, \eta_0) = D^*(P_0)$.

Estimating Equation Methods

For estimation of the causal risk difference, the following are two popular examples of estimating-equation-based methods, where the A-IPTW estimator is based on the estimating function implied by the efficient influence curve.

Estimating Equation Methods

IPTW. One estimates our target parameter, the causal risk difference $\Psi(P_0)$, with

$$\psi_n = \frac{1}{n}\sum_{i=1}^{n}\{I(A_i = 1) - I(A_i = 0)\}\frac{Y_i}{g_n(A_i, W_i)}.$$

This estimator is a solution of an IPTW estimating equation that relies on an estimate of the treatment mechanism, playing the role of a nuisance parameter of the IPTW estimating function.

Estimating Equation Methods

A-IPTW. One estimates $\Psi(P_0)$ with

$$\begin{aligned}\psi_n &= \frac{1}{n}\sum_{i=1}^{n}\frac{\{I(A_i=1)-I(A_i=0)\}}{g_n(A_i,W_i)}(Y_i-\bar{Q}_n(A_i,W_i)) \\ &+ \frac{1}{n}\sum_{i=1}^{n}\{\bar{Q}_n(1,W_i)-\bar{Q}_n(0,W_i)\}.\end{aligned}$$

Estimating Equation Methods

A-IPTW. This estimator is a solution of the A-IPTW estimating equation that relies on an estimate of the treatment mechanism g_0 and the conditional mean $\bar{Q}_0$.

Thus $(g_0, \bar{Q}_0)$ plays the role of the nuisance parameter of the A-IPTW estimating function.

The A-IPTW estimating function evaluated at the true $(g_0, \bar{Q}_0)$ and true ψ_0 actually equals the efficient influence curve at the true data-generating distribution P_0, making it an optimal estimating function.

Targeted Maximum Likelihood Estimation

TMLE (van der Laan and Rubin; 2006)

Produces a well-defined, unbiased, efficient substitution estimator of target parameters of a data-generating distribution.

It is an iterative procedure that updates an initial (super learner) estimate of the relevant part Q_0 of the data generating distribution P_0, possibly using an estimate of a nuisance parameter g_0.

TMLE

TMLE: Double Robust

- Removes asymptotic residual bias of initial estimator for the target parameter, if it uses a consistent estimator of g_0.
- If initial estimator was consistent for the target parameter, the additional fitting of the data in the targeting step may remove finite sample bias, and preserves consistency property of the initial estimator.

TMLE: Efficiency

- If the initial estimator and the estimator of g_0 are both consistent, then it is also asymptotically efficient according to semi-parametric statistical model efficiency theory.

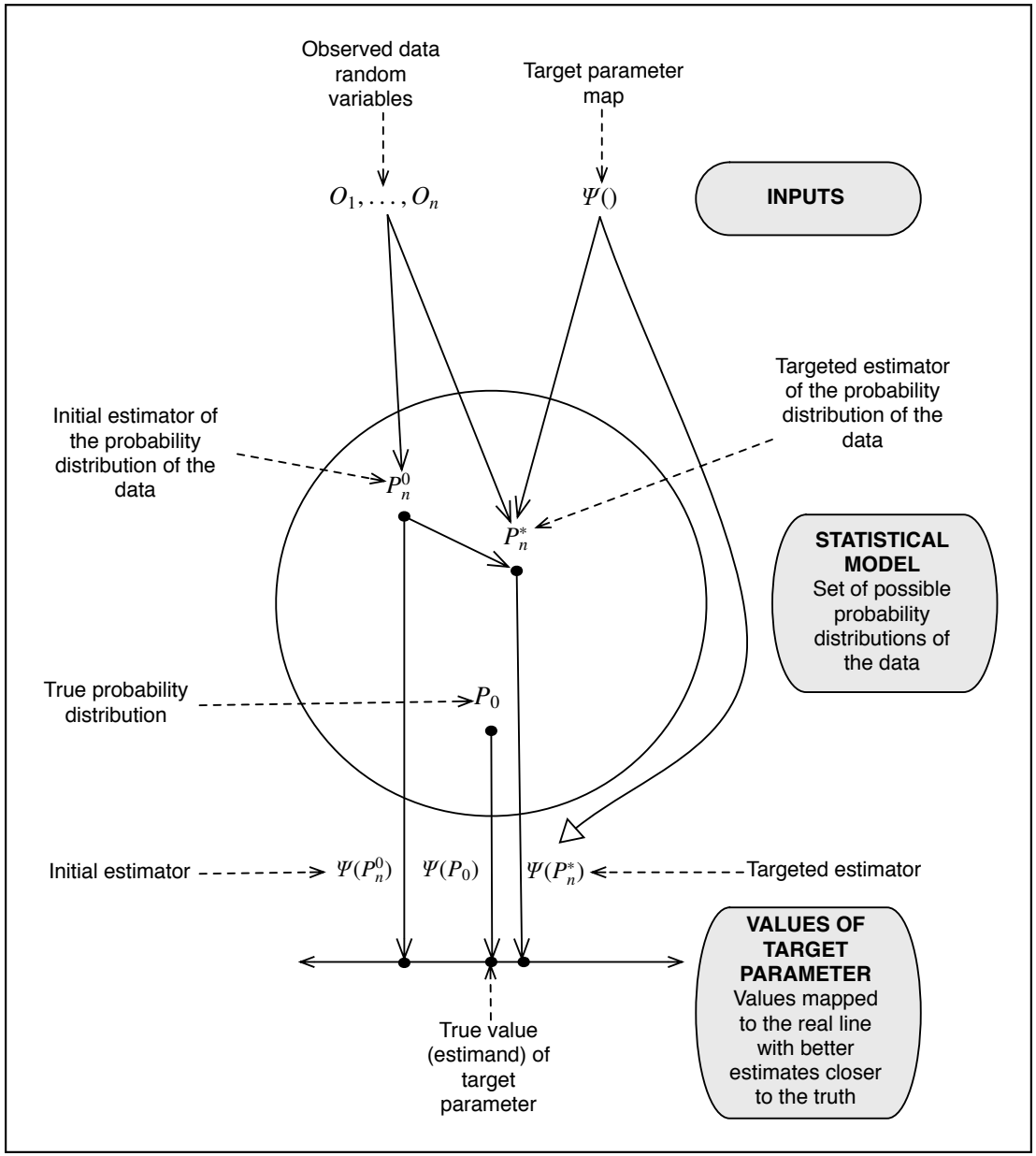
TMLE

TMLE: In Practice

Allows the incorporation of machine learning methods for the estimation of both Q_0 and g_0 so that we do not make assumptions about the probability distribution P_0 we do not believe.

Thus, every effort is made to achieve minimal bias and the asymptotic semi-parametric efficiency bound for the variance.

TMLE Algorithm



Targeted MLE

1. Idetify the parametric model for fluctuating initial $\hat{P}$
   – Small "fluctuation" $\longrightarrow$ maximum change in **target**.
2. Given strategy, identify optimum amount of fluctuation by MLE.
3. Apply optimal fluctuation to $\hat{P} \longrightarrow$ **1st-step targeted maximum likelihood estimator**.
4. Repeat until the incremental "fluctuation" is zero
   – Some important cases: 1 step to convergence.
5. Final probability distribution solves efficient influence curve equation

$\longrightarrow$ **T-MLE is double robust & locally efficient**

Targeted Minimum-Loss-Based Super Learning

$\Psi(Q_0)$ target parameter

$Q_0 = \arg\min_Q P_0 L(Q) \equiv \int L(Q)(o) dP_0(o)$

$\hat{Q}(P_n)$: Initial estimator, Loss-based SL

$\{\hat{Q}_g(\epsilon) : \epsilon\}$ fluct. model for fitting ψ_0

$\hat{g} = \hat{g}(P_n)$ loss based SL of treatment/cens mech

$\frac{d}{d\epsilon} L(\hat{Q}_{\hat{g}}(\epsilon))\Big|_{\epsilon=0} = D^*(\hat{Q}, \hat{g})$

$\epsilon_n = \arg\min_\epsilon P_n L(\hat{Q}_{\hat{g}}(\epsilon))$

Iterate till convergence: $\hat{Q}^*$

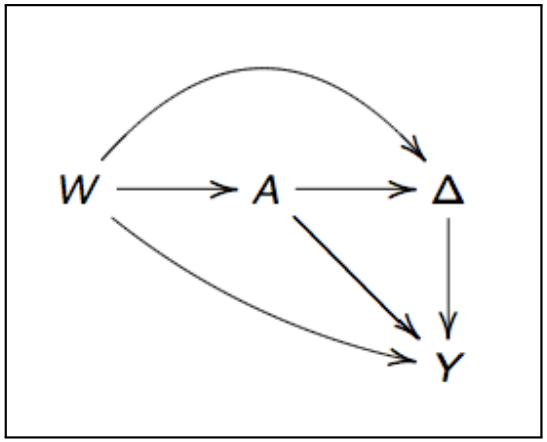
Solves efficient influence curve equation:

$P_n D^*(\hat{Q}^*, \hat{g}) = 0$

TMLE: $\Psi(\hat{Q}^*)$

Example: TMLE for the Average Causal Effect

NPSEM/SCM for a point treatment data structure with missing outcome



$$
\begin{aligned}
W &= f_W(U_W), \\
A &= f_A(W, U_A), \\
\Delta &= f_A(W, A, U_\Delta), \\
Y &= f_Y(W, A, \Delta, U_Y).
\end{aligned}
$$

We can now define counterfactuals $Y_{1,1}$ and $Y_{0,1}$ corresponding with interventions setting A and Δ.

The additive causal effect $EY_1 - EY_0$ equals:
$\Psi(P) = E[E(Y \mid A = 1, \Delta = 1, W) - E(Y \mid A = 0, \Delta, 1, W)$

Example: TMLE for the Average Causal Effect

Our first step is to generate an initial estimator of P_n^0 of P; we estimate $E(Y \mid A, \Delta = 1, W)$, possible with super learning (*to be discussed in detail on* **DAY TWO**).

We fluctuate this initial estimator with a logistic regression:

$$\text{logit} P_n^0(\epsilon)(Y = 1 \mid A, \Delta = 1, W) = \text{logit} P_n^0(Y = 1 \mid A, \Delta = 1, W) + \epsilon h$$

where

$$h(A, W) = \frac{1}{\Pi(A, W)} \left(\frac{A}{g(1 \mid W)} - \frac{1 - A}{g(0 \mid W} \right)$$

and
$g(1 \mid W) = P(A = 1 \mid W)$ Treatment Mechanism
$\Pi(A, W) = P(\Delta = 1 \mid A, W)$ Missingness Mechanism

Let ϵ_n be the maximum likelihood estimator and

$$P_n^* = P_n^0(\epsilon_n).$$

The TMLE is given by $\Psi(P_n^*)$.

Example: SPPARCS Data

- The National Institute of Aging-funded Study of Physical Performance and Age-Related Changes in Sonomans (SPPARCS) is a population-based, census-sampled, study of the epidemiology of aging and health.
- Participants of this longitudinal cohort were recruited if they were aged 54 years and over and were residents of Sonoma, CA or surrounding areas. Study recruitment of 2092 persons occurred between May 1993 and December 1994 and follow-up continued for approx. 10 years.
- The cohort was reduced to a size of $n = 2066$, as 26 subjects were missing LTPA values or self-rated health score (1.2% missing data).

Example: SPPARCS Data

- The data structure is $O = (W, A, Y)$, where $Y = I(T \leq 5 \text{ years})$, T is time to the event death, A is a binary categorization of LTPA, and W are potential confounders.
- Of note is the lack of any right censoring in this cohort. The outcome (death within or at 5 years after baseline interview) and date of death was recorded for each subject.
- Our parameter of interest is the causal risk difference, the average treatment effect of LTPA on mortality 5 years after baseline interview.

SPPARCS variables

Variable	Description
Y	Death occurring within 5 years of baseline
A	LTPA score $\geq$ 22.5 METs at baseline[‡]
W_1	Health self-rated as "excellent"
W_2	Health self-rated as "fair"
W_3	Health self-rated as "poor"
W_4	Current smoker
W_5	Former smoker
W_6	Cardiac event prior to baseline
W_7	Chronic health condition at baseline
W_8	$x \leq 60$ years old
W_9	$60 < x \leq 70$ years old
W_{10}	$80 < x \leq 90$ years old
W_{11}	$x > 90$ years old
W_{12}	Female

[‡] LTPA is calculated from a detailed questionnaire where prior performed vigorous physical activities are assigned standardized intensity values in metabolic equivalents (METs). The recommended level of energy expenditure for the elderly is 22.5 METs.

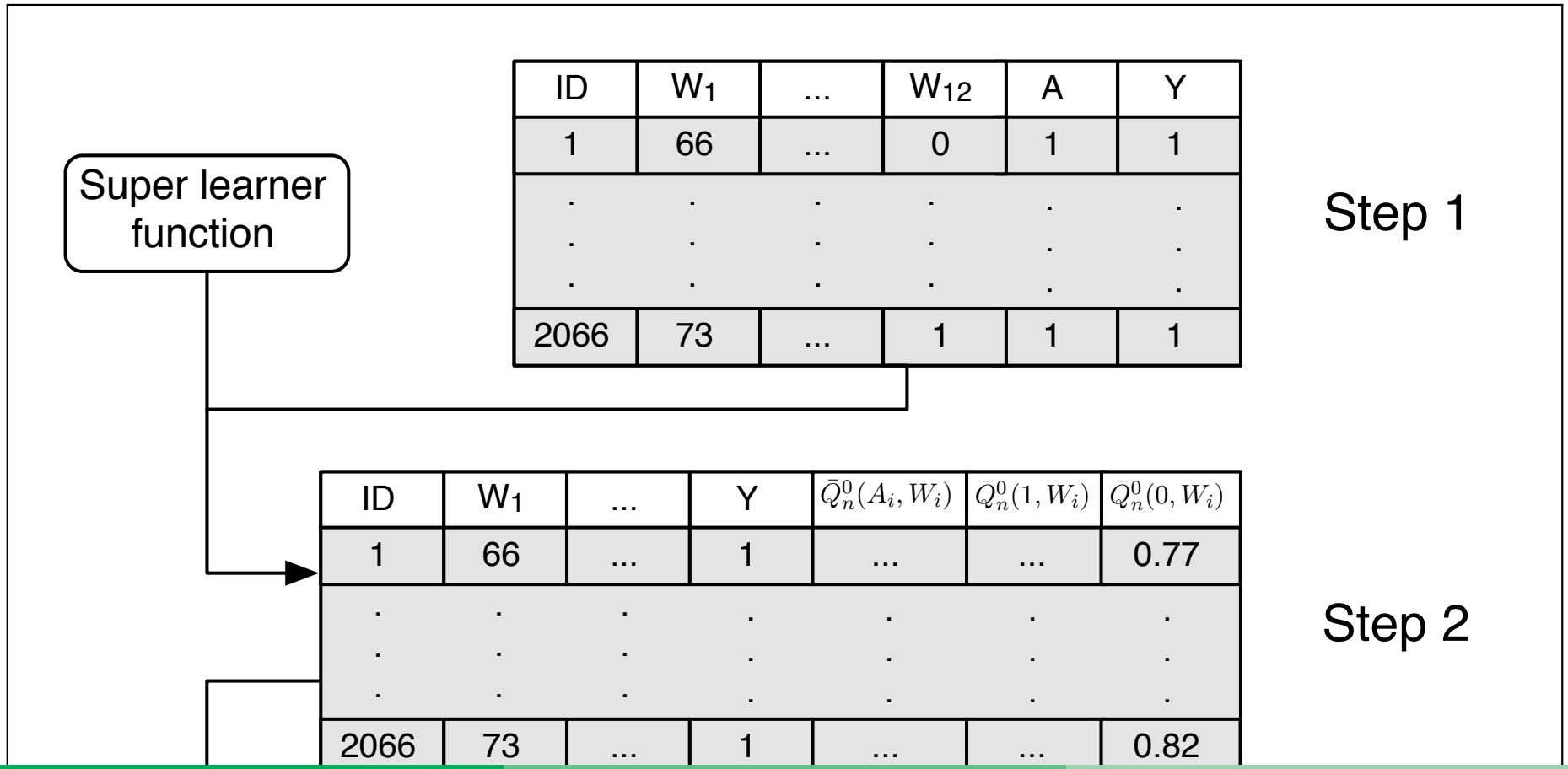
Example: SPPARCS Data: **Estimating** $\bar{Q}_0$

We take as inputs our super learner prediction function (or other choice of estimator), the initial estimate $\bar{Q}_n^0$, and our data matrix.

The data matrix includes columns for each of the covariates W found in the table, exposure LTPA (A), and outcome Y indicating death within 5 years of baseline.

(Step 1.)

TMLE Flow Diagram



Example: SPPARCS Data: **Estimating** $\bar{Q}_0$

Then we calculated predicted values for each of the 2066 observations in our data set, using their observed value of A, and added this as an n-dimensional column labeled $\bar{Q}_n^0(A_i, W_i)$ in our data matrix.

Then we calculated a predicted value for each observation where we set $a = 1$, and also $a = 0$, forming two additional columns $\bar{Q}_n^0(1, W_i)$ and $\bar{Q}_n^0(0, W_i)$. Note that for those observations with an observed value of $A_i = 1$, the value in column $\bar{Q}_n^0(A_i, W_i)$ will be equal to the value in column $\bar{Q}_n^0(1, W_i)$.

For those with observed $A_i = 0$, the value in column $\bar{Q}_n^0(A_i, W_i)$ will be equal to the value in column in $\bar{Q}_n^0(0, W_i)$.

(Step 2.)

Super learner function

ID	W_1	...	W_{12}	A	Y
1	66	...	0	1	1
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
2066	73	...	1	1	1

Step 1

ID	W_1	...	Y	$\bar{Q}_n^0(A_i, W_i)$	$\bar{Q}_n^0(1, W_i)$	$\bar{Q}_n^0(0, W_i)$
1	66	...	1	...	...	0.77
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
2066	73	...	1	...	...	0.82

Step 2

Super learner exposure mechanism function

ID	W_1	...	$\bar{Q}_n^0(0, W_i)$	$g_n(1 \mid W_i)$	$g_n(0 \mid W_i)$
1	66	...	0.77	...	0.32
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
2066	73	...	0.82	...	0.45

Step 3

Example: SPPARCS Data: **Estimating** $\bar{Q}_0$

At this stage we could plug our estimates $\bar{Q}_n^0(1, W_i)$ and $\bar{Q}_n^0(0, W_i)$ for each subject into our substitution estimator of the risk difference:

$$\psi_{MLE,n} = \Psi(Q_n) = \frac{1}{n} \sum_{i=1}^{n} \{\bar{Q}_n^0(1, W_i) - \bar{Q}_n^0(0, W_i)\}.$$

This is a super learner ML-based substitution estimator discussed previously, plugging in the empirical distribution $Q_{W,n}^0$ for the marginal distribution of W, and the super learner $\bar{Q}_n^0$ for the true regression $\bar{Q}_0$. We know that this estimator is not targeted towards the parameter of interest, so we continue on to a targeting step.

Example: SPPARCS Data: **Estimating g_0.**

Our targeting step required an estimate of the conditional distribution of LTPA given covariates W.

This estimate of $P_0(A \mid W) \equiv g_0$ is denoted g_n and can be obtained using super learning, regression, or other method.

We estimated predicted values using a super learner prediction function, adding two more columns to our data matrix: $g_n(1 \mid W_i)$ and $g_n(0 \mid W_i)$.

(Step 3.)

ID	W_1	...	Y	$\bar{Q}_n^0(A_i, W_i)$	$\bar{Q}_n^0(1, W_i)$	$\bar{Q}_n^0(0, W_i)$
1	66	...	1	...	...	0.77
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
2066	73	...	1	...	...	0.82

Step 2

Super learner exposure mechanism function

ID	W_1	...	$\bar{Q}_n^0(0, W_i)$	$g_n(1 \mid W_i)$	$g_n(0 \mid W_i)$
1	66	...	0.77	...	0.32
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
2066	73	...	0.82	...	0.45

Step 3

ID	W_1	...	$g_n(0 \mid W_i)$	$H_n^*(A_i, W_i)$	$H_n^*(1, W_i)$	$H_n^*(0, W_i)$
1	66	...	0.32	...	...	-3.13
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
2066	73	...	0.45	...	...	-2.22

Step 4

Example: SPPARCS Data: **Determining a parametric working model to fluctuate the initial estimator.**

The targeting step used the estimate g_n in a clever covariate to define a parametric working model coding fluctuations of the initial estimator. This clever covariate $H_n^*(A, W)$ is given by

$$H_n^*(A, W) \equiv \left(\frac{I(A = 1)}{g_n(1 \mid W)} - \frac{I(A = 0)}{g_n(0 \mid W)} \right).$$

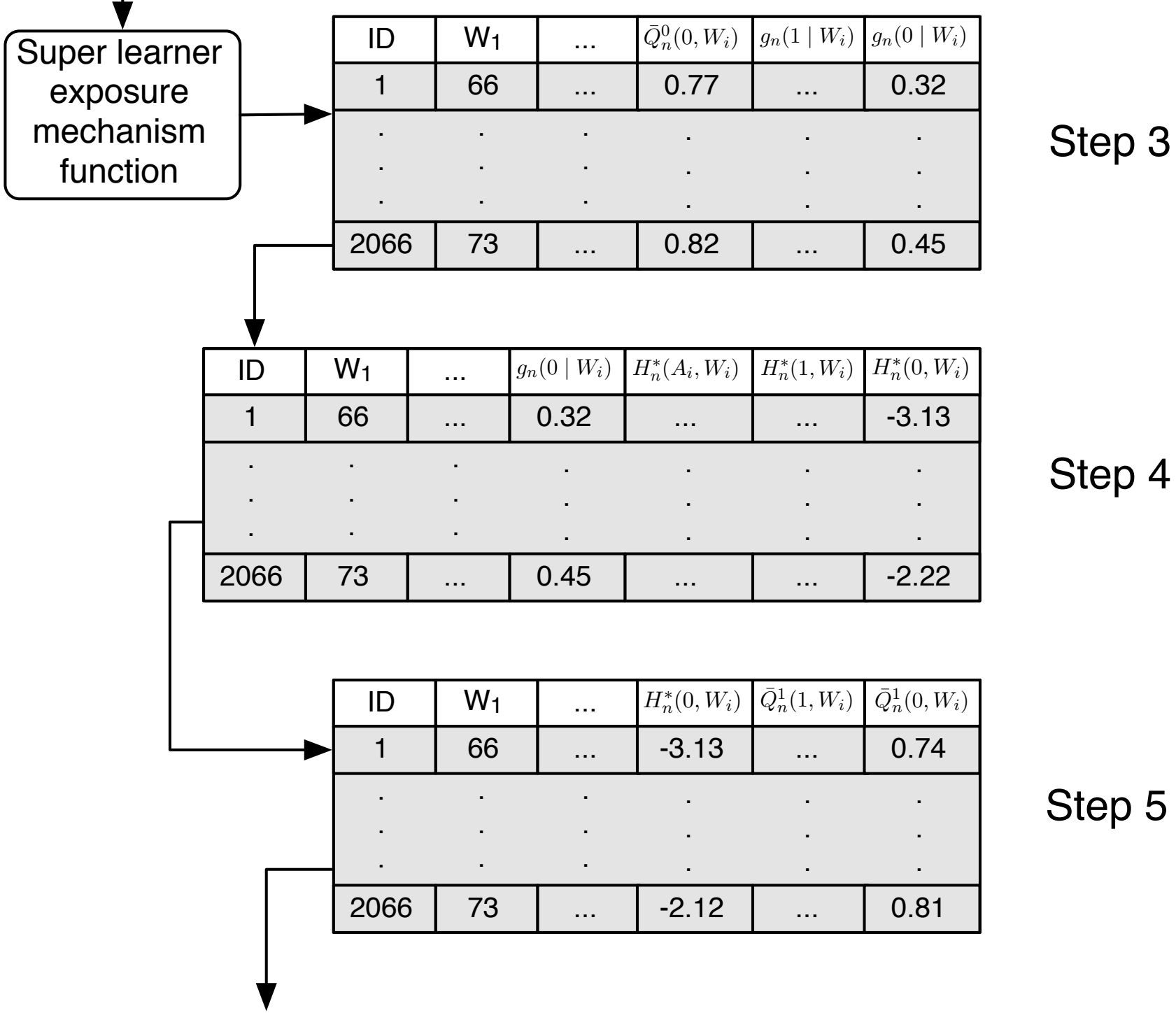
Example: SPPARCS Data: **Determining a parametric working model to fluctuate the initial estimator.**

Thus, for each subject with $A_i = 1$ in the observed data, we calculated the clever covariate as $H_n^*(1, W_i) = 1/g_n(1 \mid W_i)$.

Similarly, for each subject with $A_i = 0$ in the observed data, we calculated the clever covariate as $H_n^*(0, W_i) = -1/g_n(0 \mid W_i)$.

We combined these values to form a single column $H_n^*(A_i, W_i)$ in the data matrix. We also added two columns $H_n^*(1, W_i)$ and $H_n^*(0, W_i)$. The values for these columns were generated by setting $a = 0$ and $a = 1$.

(Step 4.)


Super learner exposure mechanism function
ID W1 ... $\bar{Q}_n^0(0, W_i)$ $g_n(1 \mid W_i)$ $g_n(0 \mid W_i)$
1 66 ... 0.77 ... 0.32
2066 73 ... 0.82 ... 0.45
Step 3
ID W1 ... $g_n(0 \mid W_i)$ $H_n^*(A_i, W_i)$ $H_n^*(1, W_i)$ $H_n^*(0, W_i)$
1 66 ... 0.32 ... ... -3.13
2066 73 ... 0.45 ... ... -2.22
Step 4
ID W1 ... $H_n^*(0, W_i)$ $\bar{Q}_n^1(1, W_i)$ $\bar{Q}_n^1(0, W_i)$
1 66 ... -3.13 ... 0.74
2066 73 ... -2.12 ... 0.81
Step 5

Example: SPPARCS Data: **Updating $\bar{Q}_n^0$.**

We then ran a logistic regression of our outcome Y on the clever covariate using as intercept the offset $\text{logit}\bar{Q}_n^0(A, W)$ to obtain the estimate ϵ_n, where ϵ_n is the resulting coefficient in front of the clever covariate $H_n^*(A, W)$.

We next wanted to update the estimate $\bar{Q}_n^0$ into a new estimate $\bar{Q}_n^1$ of the true regression function $\bar{Q}_0$:

$$\text{logit}\, \bar{Q}_n^1(A, W) = \text{logit}\, \bar{Q}_n^0(A, W) + \epsilon_n H_n^*(A, W).$$

This parametric working model incorporated information from g_n, through $H_n^*(A, W)$, into an updated regression.

Example: SPPARCS Data: **Updating $\bar{Q}_n^0$.**

One can now repeat this updating step by running a logisitic regression of outcome Y on the clever covariate $H_n^*(A, W)$ using as intercept the offset logit $\bar{Q}_n^1(A, W)$ to obtain the next update $\bar{Q}_n^2$.

However, it follows that this time the coefficient in front of the clever covariate will be equal to zero, so that subsequent steps do not result in further updates.

Convergence of the TMLE algorithm was achieved in one step.

Example: SPPARCS Data: **Updating $\bar{Q}_n^0$.**

The TMLE of Q_0 was given by $Q_n^* = (\bar{Q}_n^1, Q_{W,n}^0)$. With ϵ_n, we were ready to update our prediction function at $a = 1$ and $a = 0$ according to the logistic regression working model. We calculated

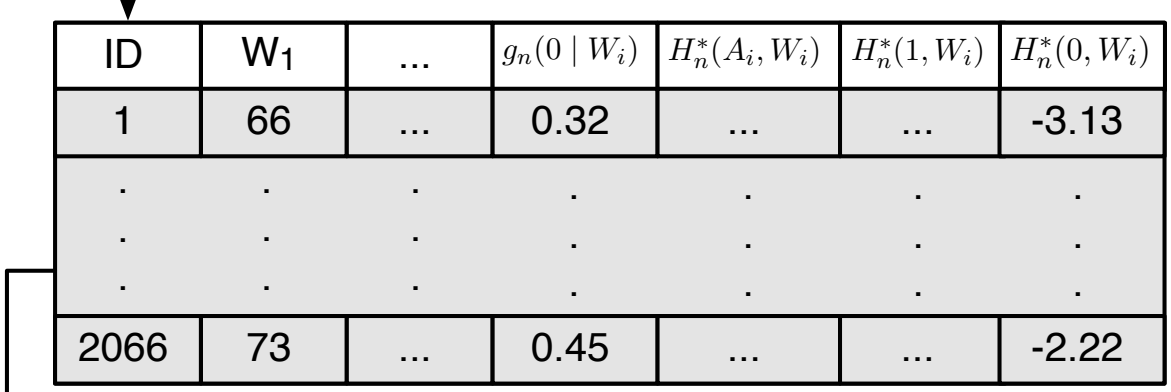
$$\text{logit}\, \bar{Q}_n^1(1, W) = \text{logit}\bar{Q}_n^0(1, W) + \epsilon_n H_n^*(1, W),$$

for all subjects, and then

$$\text{logit}\, \bar{Q}_n^1(0, W) = \text{logit}\bar{Q}_n^0(0, W) + \epsilon_n H_n^*(0, W)$$

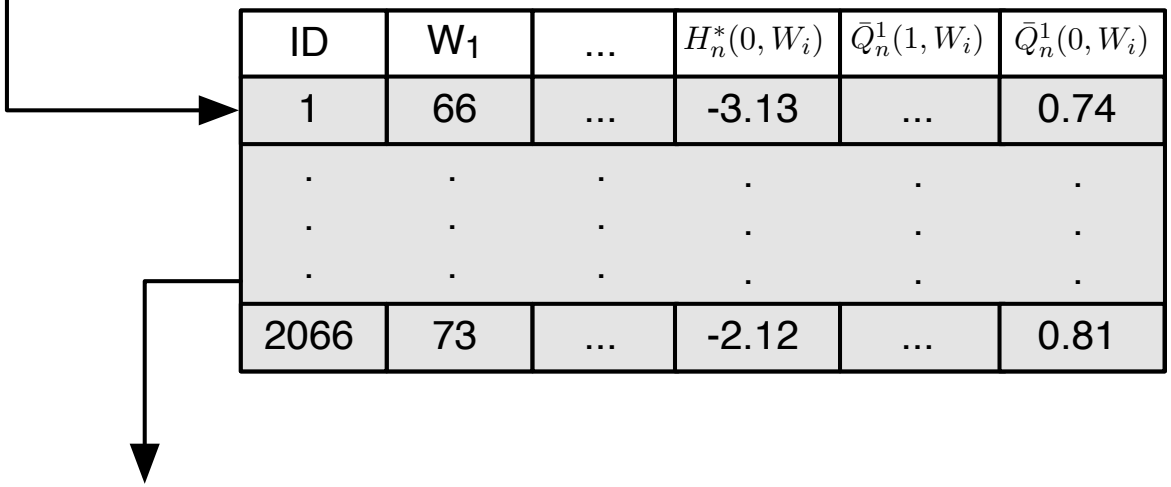
for all subjects and added a column for $\bar{Q}_n^1(1, W_i)$ and $\bar{Q}_n^1(0, W_i)$ to the data matrix.

Updating $\bar{Q}_n^0$ is also illustrated in Step 5.



ID	W_1	...	$g_n(0 \mid W_i)$	$H_n^*(A_i, W_i)$	$H_n^*(1, W_i)$	$H_n^*(0, W_i)$
1	66	...	0.32	...	...	-3.13
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
2066	73	...	0.45	...	...	-2.22

Step 4



ID	W_1	...	$H_n^*(0, W_i)$	$\bar{Q}_n^1(1, W_i)$	$\bar{Q}_n^1(0, W_i)$
1	66	...	-3.13	...	0.74
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
2066	73	...	-2.12	...	0.81

Step 5

$$\psi_n = \frac{1}{n} \sum_{i=1}^{n} [\bar{Q}_n^1(1, W_i) - \bar{Q}_n^1(0, W_i)]$$

Step 6

Example: SPPARCS Data: **Targeted substitution estimator of the target parameter.**

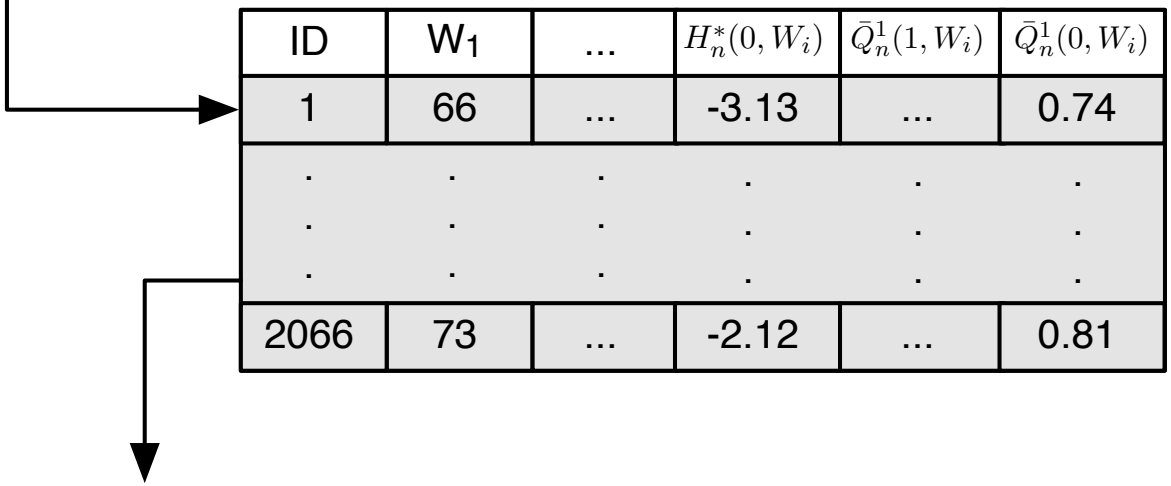
We computed the plug-in targeted maximum likelihood substitution estimator using the updated estimates $\bar{Q}_n^1(1, W)$ and $\bar{Q}_n^1(0, W)$ and the empirical distribution of W (Step 6.)

Our formula from the first step becomes

$$\psi_{TMLE,n} = \Psi(Q_n^*) = \frac{1}{n} \sum_{i=1}^{n} \{\bar{Q}_n^1(1, W_i) - \bar{Q}_n^1(0, W_i)\}.$$

This mapping was accomplished by evaluating $\bar{Q}_n^1(1, W_i)$ and $\bar{Q}_n^1(0, W_i)$ for each observation i, and plugging these values into the above equation.

Our estimate of the causal risk difference for the mortality study was $\psi_{TMLE,n} = -0.055$.



ID	W_1	...	$H_n^*(0, W_i)$	$\bar{Q}_n^1(1, W_i)$	$\bar{Q}_n^1(0, W_i)$
1	66	...	-3.13	...	0.74
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
2066	73	...	-2.12	...	0.81

Step 5

$$\psi_n = \frac{1}{n} \sum_{i=1}^{n} [\bar{Q}_n^1(1, W_i) - \bar{Q}_n^1(0, W_i)]$$

Step 6

Example: SPPARCS Data: **Inference (Standard errors).**

We then needed to calculate the influence curve for our estimator in order to obtain standard errors:

$$IC_n(O_i) = \left(\frac{I(A_i = 1)}{g_n(1 \mid W_i)} - \frac{I(A_i = 0)}{g_n(0 \mid W_i)} \right) (Y - \bar{Q}_n^1(A_i, W_i)) + \bar{Q}_n^1(1, W_i) - \bar{Q}_n^1(0, W_i) - \psi_{TMLE,n},$$

where I is an indicator function: it equals 1 when the logical statement it evaluates, e.g., $A_i = 1$, is true.

Example: SPPARCS Data: **Inference (Standard errors).**

Note that this influence curve is evaluated for each of the n observations O_i.

The beauty of the influence curve of an estimator is that one can now proceed with statistical inference as if the estimator minus its estimand equals the empirical mean of the influence curve.

Example: SPPARCS Data: **Inference (Standard errors).**

Next, we calculated the sample mean of these estimated influence curve values: $\bar{IC}_n = \frac{1}{n}\sum_{i=1}^{n} IC_n(o_i)$, where we use o_i to stress that this mean is calculated with our observed realizations of the random variable O_i. For the TMLE we have $\bar{IC}_n = 0$. Using this mean, we calculated the sample variance of the estimated influence curve values:

$$S^2(IC_n) = \frac{1}{n}\sum_{i=1}^{n} \left(IC_n(o_i) - \bar{IC}_n\right)^2.$$

Lastly, we used our sample variance to estimate the standard error of our estimator:

$$\sigma_n = \sqrt{\frac{S^2(IC_n)}{n}}.$$

This estimate of the standard error in the mortality study was $\sigma_n = 0.012$.

Example: SPPARCS Data: **Inference (Confidence intervals).**

With the standard errors, we can now calculate confidence intervals and p-values in the same manner you may have learned in other statistics texts. A 95% Wald-type confidence interval can be constructed as:

$$\psi_{TMLE,n} \pm z_{0.975} \frac{\sigma_n}{\sqrt{n}},$$

where z_α denotes the α-quantile of the standard normal density $N(0,1)$.

Example: SPPARCS Data: **Inference (p-values).**

A p-value for $\psi_{TMLE,n}$ can be calculated as:

$$2\left[1-\Phi\left(\left|\frac{\psi_{TMLE,n}}{\sigma_n/\sqrt{n}}\right|\right)\right],$$

where Φ denotes the standard normal cumulative distribution function.

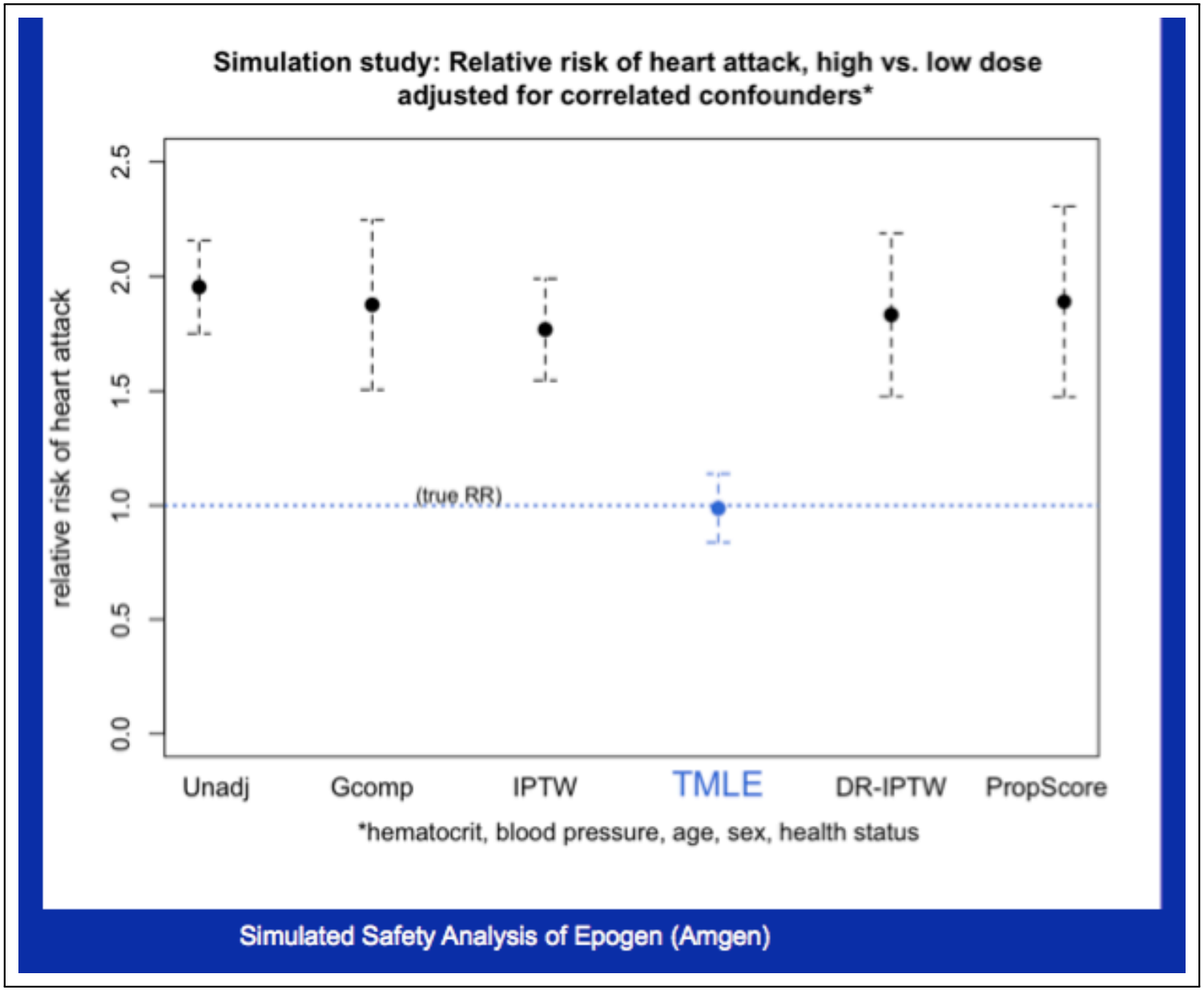
The p-value was < 0.001 and the confidence interval was $[-0.078, -0.033]$.

Example: SPPARCS Data: **Interpretation.**

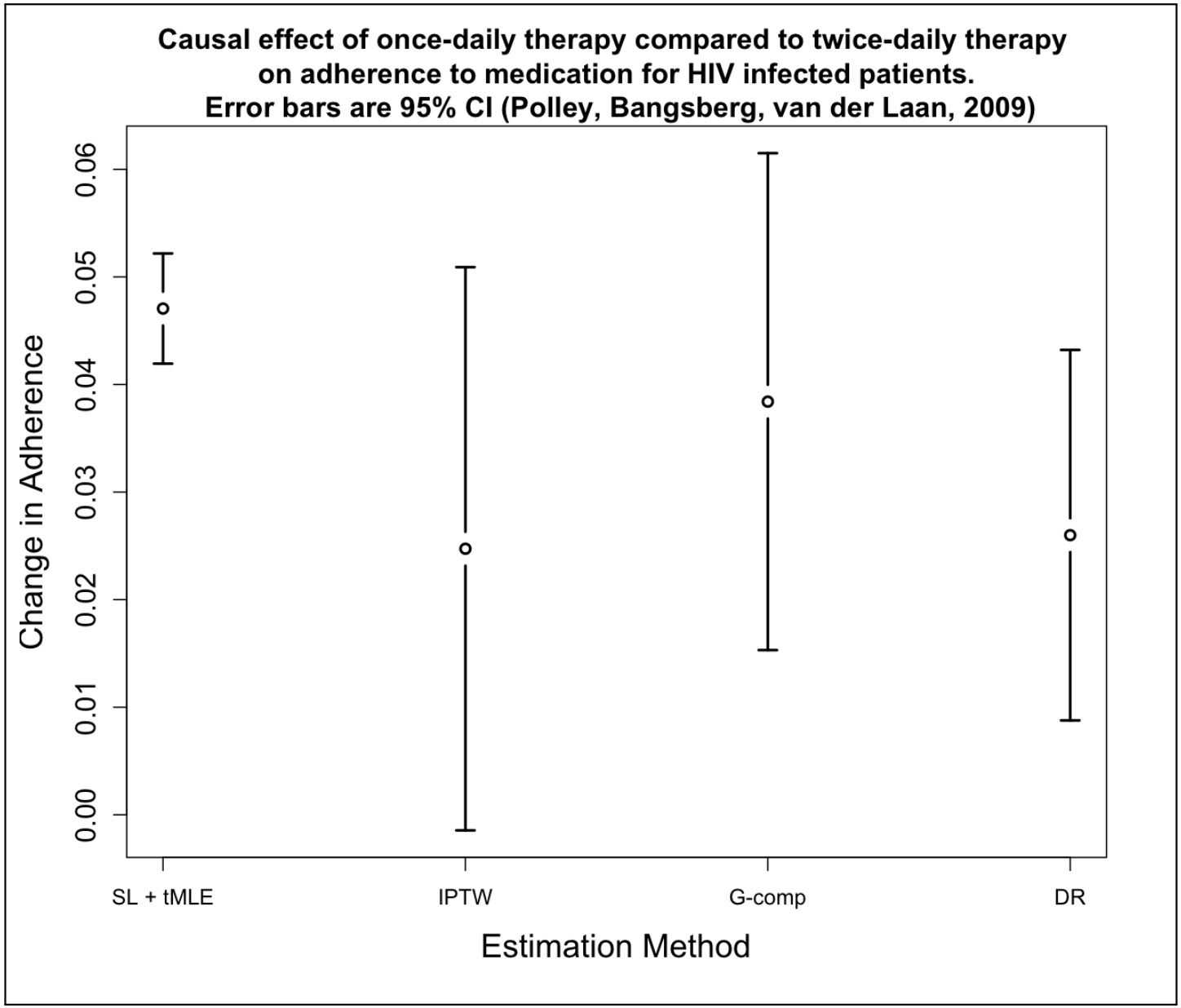
The interpretation of our estimate $\psi_{TMLE,n} = -0.055$, under causal assumptions, is that meeting or exceeding recommended levels of LTPA decreases 5-year mortality in an elderly population by 5.5%.

This result was significant, with a p-value of < 0.001 and a confidence interval of $[-0.078, -0.033]$.

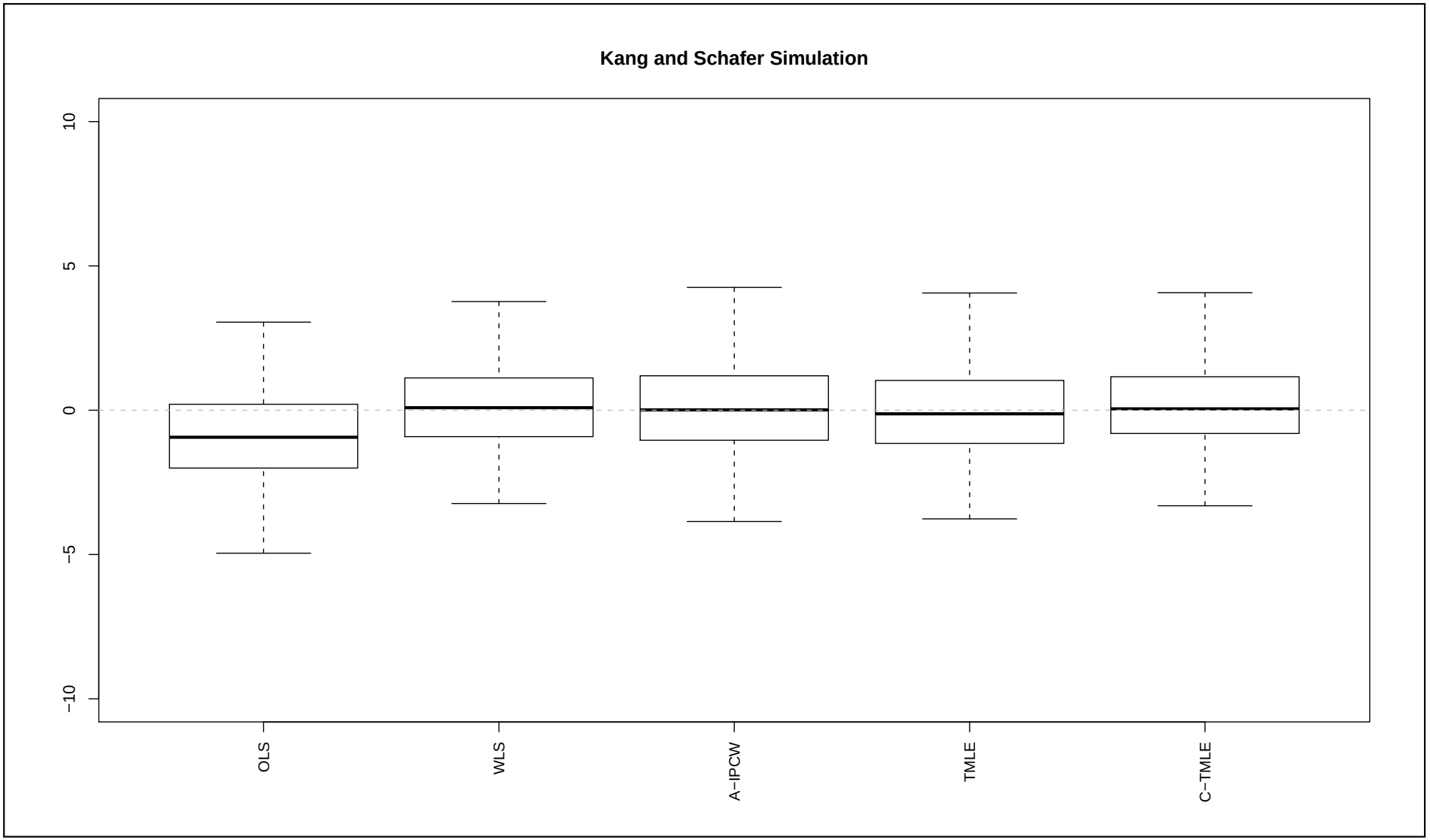
Some Simulation Results



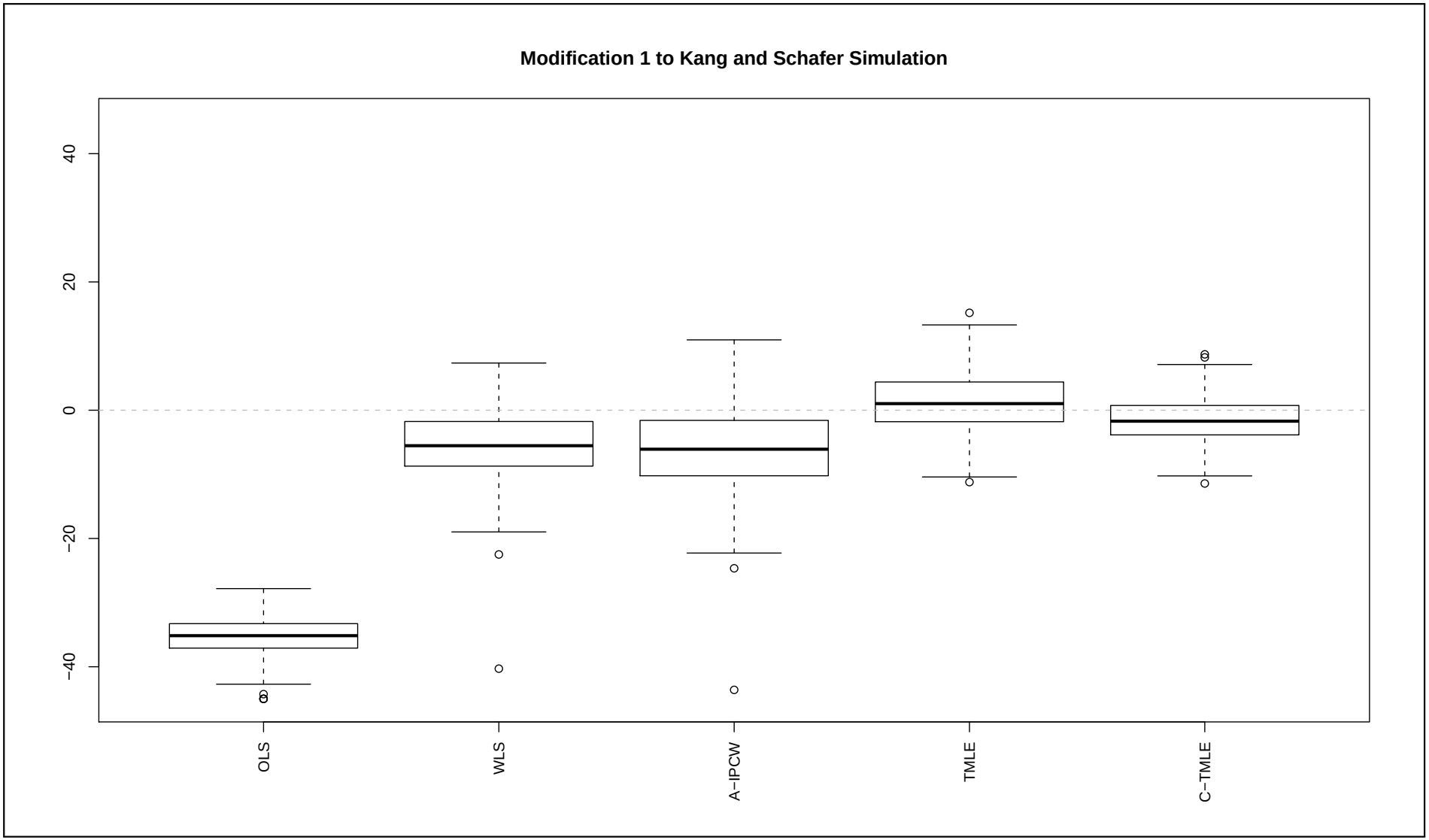
Some Simulation Results



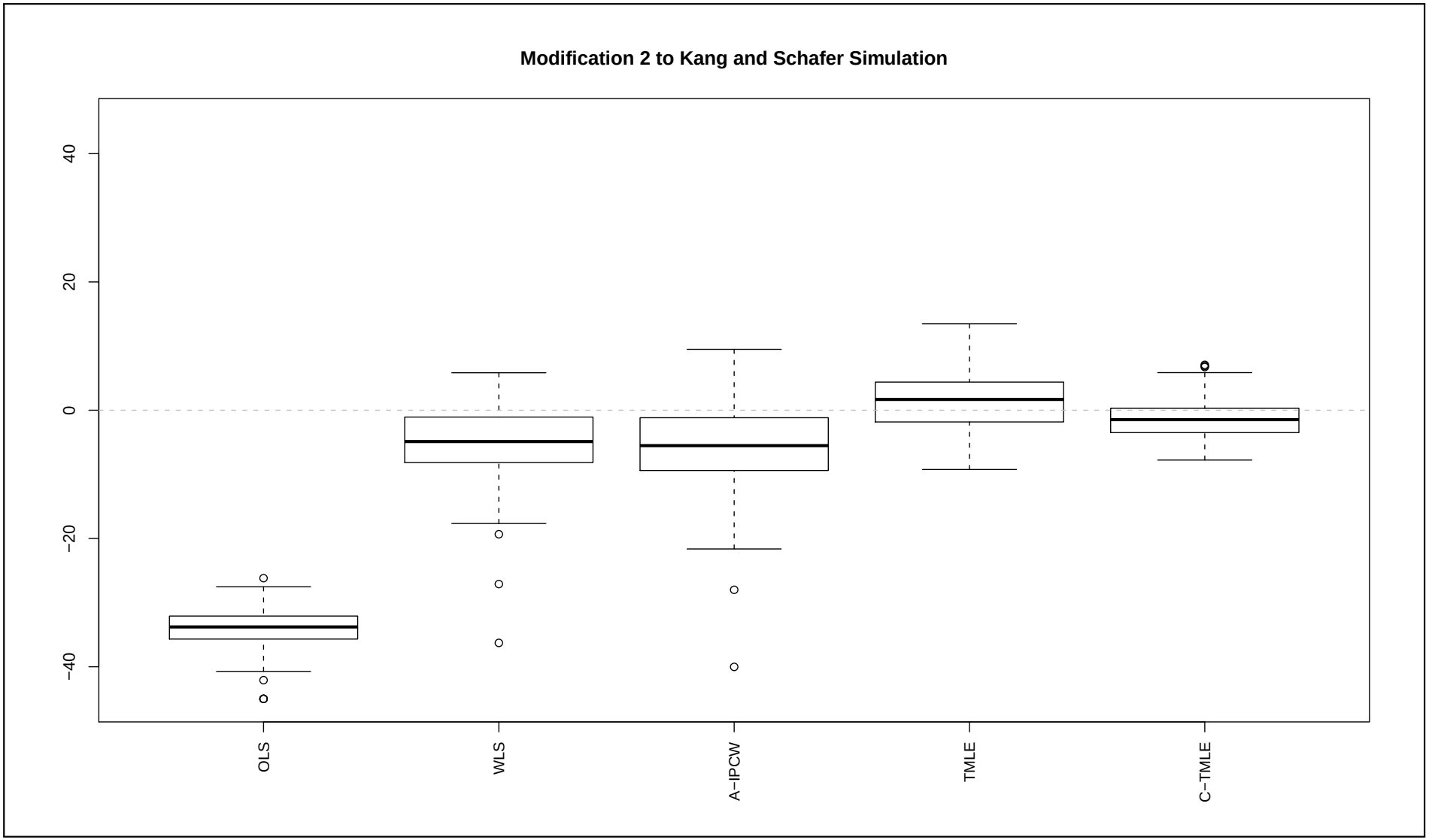
Some Simulation Results



Some Simulation Results



Some Simulation Results



A General Roadmap for Causal Inference

- Specify the Question, Data, Likelihood
- Specify NPSEM/SCM
- Specify the Causal Parameter of Interest
- Asses Identifiability
- Commit to a Statistical Model and Target
- Estimate the Target Parameter
  –Semiparametric efficient subsitution estimator
  –Targeted minimum loss-based super learning