R Assignment 5 - TMLE

Introduction to Causal Inference

Write-up: Please answer all questions and include relevant `R` code. You are encouraged to discuss the assignment in groups, but should not copy code or interpretations verbatim.

1 Implement TMLE for the G-Computation estimand

Suppose our collaborators are interested in using TMLE to understand the impact of completed burpees on expected happiness among the subpopulation of BART riders, who complete $A = 1$ or $A = 7$ burpees. (See R assignment 4 for the background story.) Under the necessary causal assumptions, the statistical estimand is given by the G-Computation formula for this target population:

$$\Psi(P_0) = \sum_w \big[E_0(Y|A = 7, W = w) - E_0(Y|A = 1, W)\big] P_0(W = w|A = 1 \text{ or } A = 7)$$

This is an example of a conditional target parameter. Other examples of conditional (causal) parameters include the average treatment effect among the treated or the average treatment effect among the untreated.

1. Set the seed to 252.
2. Import the data set `RAssign4.csv` and assign it to object `FullData`.
3. Create data frame `ObsData`, consisting riders completing $A = 1$ or $A = 7$ burpees (i.e. our subpopulation of interest):

   ```
   > ObsData<- FullData[FullData$A==1 | FullData$A==7, ]
   ```

4. Assign the number of riders in `ObsData` to `n`.
5. Create a new exposure variable `A.binary`, which equals 1 for riders completing $A = 7$ burpees and equals 0 for riders completing $A = 1$ burpee.
6. Use the `table` function to make sure your code is correct.
7. Implement `tmle` using SuperLearner with the default library for initial estimation of $\bar{Q}_0(A, W)$ and $g_0(A|W)$. Be sure to specify the outcome (`Y=ObsData$Y`), the exposure (`A=A.binary`) and the covariates (`W=subset(ObsData, select=c(W1,W2))`).

2 Evaluate the finite sample performance of TMLE

1. Set the seed to `set.seed(252)`. Create a vector `estimates` of size `R=500`.
2. Within a `for` loop, repeat the following `R=500` times.
   (a) Draw a sample of size 5,000 independently from data generating process given below.
   (b) Create data frame `ObsData`, consisting riders completing $A = 1$ or $A = 7$ burpees.
   (c) Assign the number of riders in `ObsData` to `n`.

(d) Create a new exposure variable `A.binary`, which equals 1 for riders completing $A = 7$ burpees and equals 0 for riders completing $A = 1$ burpee.

(e) Implement `tmle` using SuperLearner with the default library for initial estimation of $\bar{Q}_0(A, W)$ and $g_0(A|W)$.

(f) Save the resulting point estimate as a row in `estimates`.

3. What is the average estimate? What is the bias (average deviation from the point estimate and the true value)? How variable are the estimates?

4. Create a histogram of the point estimates.
Hint: Use `hist` function.

```
> # ----------------
> # generateData - function to generate the observed data + counterfactuals
> # ----------------
> # this does (should) NOT need to be in the for loop.
> generateData<- function(n){
+
+   W1<- as.integer(runif(n, 1,4)  ) # lifestyle 1,2,3
+   W2<- rbinom(n, size=1, prob= runif(n, 0.02, 0.7)) # gender
+   A<- 1+ rbinom(n, size=6, prob=plogis(0.35 -0.3*W1 +0.5*(1-W2) )) #burpees
+   U.Y<- rnorm(n, 0, sd=0.01)
+   Y<- 30 +1.5*W1 +3*log(A)+.3*(1-W2)*A + U.Y # happiness
+
+   #  the counterfactuals
+   Y.1<- 30 +1.5*W1 +3*log(1)+.3*(1-W2)*1 + U.Y #
+   Y.2<- 30 +1.5*W1 +3*log(2)+.3*(1-W2)*2 + U.Y #
+   Y.3<- 30 +1.5*W1 +3*log(3)+.3*(1-W2)*3 + U.Y #
+   Y.4<- 30 +1.5*W1 +3*log(4)+.3*(1-W2)*4 + U.Y #
+   Y.5<- 30 +1.5*W1 +3*log(5)+.3*(1-W2)*5 + U.Y #
+   Y.6<- 30 +1.5*W1 +3*log(6)+.3*(1-W2)*6 + U.Y #
+   Y.7<- 30 +1.5*W1 +3*log(7)+.3*(1-W2)*7 + U.Y #
+
+   data.frame(W1,W2,A,Y,Y.1, Y.2, Y.3, Y.4,Y.5, Y.6, Y.7)
+ }
> # ---------------
```