R Lab 5 - TMLE

Introduction to Causal Inference

Goals:
1. Review the causal roadmap.
2. Code TMLE for the G-computation estimand.
3. Understand the basics of the `tmle` package.
4. Use the `tmle` package to explore the double robustness of TMLE.

Next lab:
We will implement the non-parametric bootstrap to estimate the standard error of the estimators. We will also use the sample variance of the estimated influence curve to obtain inference for TMLE.

1 Background

Dr. Alan Grant: "T-Rex doesn't want to be fed. He wants to hunt. Can't just suppress 65 million years of gut instinct." - Michael Crichton

We are interested in estimating the causal effect of prior experience with Dinosaurs on survival on Isla Nublar, the location of the InGen lab. Suppose we have data on the following variables:

- W1: gender (1 for male; 0 for female)
- W2: intelligence (1 for smart; 0 for not)
- W3: handy/inventiveness (scale from 0 for none to 1 for MacGyver)
- W4: running speed (scale from 0 for slow to 3 for fast)
- A: prior Dinosaur experience (1 for yes; 0 for no)
- Y: survival (1 for yes; 0 for no)

Let $W = (W1, W2, W3, W4)$ be the vector of baseline covariates.

http://www.thesambarnes.com/web-project-management/account-management-for-the-web-project-manager-part-1/

2 Causal Road Map Rundown

1. **Specify the Question:**
   What is the causal effect of prior experience on survival in Jurassic Park?

2. **Specify the causal model:**
   - Endogenous nodes: $X = (W, A, Y)$, where $W = (W1, W2, W3, W4)$ is the set of baseline covariates (gender, intelligence, MacGyver-ness, running speed), A is prior Dinosaur experience and Y is survival. For simplicity, we have condensed the baseline characteristics into a single node.
   - Exogenous nodes: $U = (U_W, U_A, U_Y) \sim P_U$. We place no assumptions on the distribution P_U.
   - Structural equations F:

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

   We have made exclusion restrictions, but not placed any restrictions on the functional forms.

3. **Specify the causal parameter of interest:**
   We are interested in the causal effect of prior Dinosaur experience on survival on Isla Nublar (i.e. the causal risk difference or the average treatment effect):

$$\Psi^F(P_{U,X}) = P_{U,X}(Y_1 = 1) - P_{U,X}(Y_0 = 1) = E_{U,X}(Y_1) - E_{U,X}(Y_0)$$

   where Y_a is the counterfactual outcome (survival), if possibly contrary to fact, the subject had experience $A = a$.

4. **Specify the link between the SCM and the observed data:**
   We assume that the observed data $O = (W, A, Y) \sim P_0$ were generated by sampling n times from a data generating described by the SCM. The statistical model $\mathcal{M}$ for the set of allowed distributions of the observed data is non-parametric.

5. **Assess identifiability:**
   In the original SCM $\mathcal{M}^F$, the target causal parameter is not identified from the observed data distribution. We need make assumptions about the independence of exogenous errors: $U_A \perp\!\!\!\perp U_Y$ and (i) $U_A \perp\!\!\!\perp U_W$, or (ii) $U_Y \perp\!\!\!\perp U_W$. Then the backdoor criteria will hold conditionally on $W = (W1, W2, W3, W4)$. We use M^{F*} to denote the original SCM augmented by the assumptions needed for identifiability.

   To identify $E_{U,X}(Y_a)$ with the G-Computation formula, we also need the positivity assumption to hold

$$min_{a \in \mathcal{A}}\ P_0(A = a | W = w) > 0$$

   for all w for which $P_0(W = w) > 0$. In terms of our example, there must be a positive probability of being experienced and not experienced within strata of baseline covariates.

6. **Specify the statistical estimand:**
   The target parameter of the observed data distribution (which equals the causal parameter in the augmented causal model) is given by the G-Computation formula:

$$\Psi(P_0) = E_0\big[E_0(Y|A = 1, W = w) - E_0(Y|A = 0, W = w)\big]$$

   This is our statistical estimand.

7. **Estimate the chosen parameter of the observed data distribution:**

   (a) **Simple substitution estimator based on the G-Computation formula:**

$$\hat{\Psi}_{SS}(P_n) = \frac{1}{n} \sum_{i=1}^{n} \big(\bar{Q}_n(1, W_i) - \bar{Q}_n(0, W_i)\big)$$

where P_n is the empirical distribution and $\bar{Q}_n(A, W)$ is the estimate of the conditional mean outcome given the exposure (experience with Dinosaurs) and baseline covariates.
- Consistency of the simple (non-targeted) substitution estimator depends on consistent estimation of $\bar{Q}_0(A, W) = E_0(Y|A, W)$.

(b) **Standard (unstabilized) inverse probability weighted estimator (IPTW)**:

$$\hat{\Psi}_{IPTW}(P_n) = \frac{1}{n}\sum_{i=1}^{n}\frac{\mathbb{I}(A_i = 1)}{g_n(A_i|W_i)}Y_i - \frac{1}{n}\sum_{i=1}^{n}\frac{\mathbb{I}(A_i = 0)}{g_n(A_i|W_i)}Y_i$$

where $g_n(A_i|W_i) = P_n(A_i|W_i)$ is an estimate of the treatment mechanism (i.e. the conditional probability of having Dinosaur experience, given the baseline covariates).
- Consistency of IPTW estimators depends on consistent estimation of $g_0(A|W)$.

(c) **Targeted maximum likelihood estimation (TMLE):**

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

where $\bar{Q}_n^*(A, W)$ denotes the updated estimate of the conditional mean outcome given the exposure and baseline covariates.
- Implementation requires estimation of both the conditional mean function $\bar{Q}_0(A, W)$ and the treatment mechanism $g_0(A|W)$.
- Double robust estimators are consistent if *either* $\bar{Q}_0(A, W)$ *or* $g_0(A|W)$ are estimated consistently.
- If both $\bar{Q}_0(A, W)$ and $g_0(A|W)$ are estimated consistently, TMLE will be efficient and achieve the lowest possible asymptotic variance over a large class of estimators.
- These asymptotic properties describe what happens when sample size goes to infinity and also translate into lower bias and variance in finite samples.

If we apply an estimator to our observed data (n i.i.d. copies of O drawn from P_0), we get an estimate (a number). The estimator is function of a random variable; so it is a random variable. It has a distribution, which we can study theoretically or using simulations.

Note: An estimator is *consistent* if the point estimates converge (in probability) to the estimand as sample size $n \to \infty$.

8. **Inference and interpret results:**
In the next lab, we will implement the non-parametric bootstrap for variance estimation for the three types of estimators. We will use the sample variance of the estimated influence curve to obtain inference for the TMLE.

3 Import and explore data set `RLab5.TMLE.csv`.

1. Use the `read.csv` function to import the data set and assign it to data frame `ObsData`.
2. Use the `head` and `summary` functions to explore the data.
3. Use the `nrow` function to count the number of subjects in the data set. Assign this number as `n`.

4 Implement TMLE for the G-computation estimand

1. **Use SuperLearner to estimate $E_0(Y|A, W) = \bar{Q}_0(A, W)$, which is the conditional probability of surviving given the exposure (prior experience) and baseline covariates.**

(a) Use the `library` function to load the `SuperLearner` package and then specify the SuperLearner library with the following algorithms: `SL.glm`, `SL.step` and `SL.glm.interaction`.

```
> library("SuperLearner")
> # specify the library
> SL.library<- c("SL.glm", "SL.step", "SL.glm.interaction")
```

(b) Create data frame `X` consisting of the covariates $(W1, W2, W3, W4)$ and the intervention A.

i. Also create data frame `X1` where A has been set to 1.

ii. Also create data frame `X0` where A has been set to 0.

iii. Finally, create data frame `newdata` by stacking the data frames `X`, `X1`, `X0`:

```
> newdata<- rbind(X,X1,X0)
```

We will use `newdata` to obtain the expected outcome under the observed exposure $\bar{Q}_n(A, W)$, under the intervention $\bar{Q}_n(A = 1, W)$ and under the control $\bar{Q}_n(A = 0, W)$.

(c) Estimate $\bar{Q}_0(A, W)$ by running `SuperLearner`. Call this object `Qinit`. Be sure to specify the `SL.library` and the appropriate `family`.

```
> Qinit<- SuperLearner(Y=ObsData$Y, X=X, newX=newdata, SL.library=SL.library,
+     family="binomial")
```

Including `newX=newdata` allows us to get the predicted outcomes for all subjects under their observed exposure (i.e. using (A, W) in `X`), under the treatment (i.e. using $(A = 1, W)$ in `X1`), and under the control (i.e. $(A = 0, W)$ in `X0`)

(d) The predicted probabilities of surviving are accessed with `Qinit$SL.predict`. This is a vector of length $3n$.

i. Assign the predicted probability of surviving, given the subject's observed experience A and baseline characteristics W to `QbarAW`:

```
> QbarAW <- Qinit$SL.predict[1:n]
```

ii. Assign the predicted probability of survival for each subject given $A = 1$ and W to `Qbar1W`:

```
> Qbar1W <- Qinit$SL.predict[(n+1):(2*n)]
```

iii. Assign the predicted probability of survival for each subject given $A = 0$ and W to `Qbar0W`:

```
> Qbar0W <- Qinit$SL.predict[(2*n+1):(3*n)]
```

(e) Evaluate the simple substitution estimator by plugging the estimates $\bar{Q}_n(1, W)$ and $\bar{Q}_n(0, W)$ into the target parameter mapping:

$$\hat{\Psi}_{SS}(P_n) = \frac{1}{n} \sum_{i=1}^{n} \bar{Q}_n(1, W_i) - \bar{Q}_n(0, W_i)$$

Note: This step is not part of the TMLE algorithm, but done for comparison.

2. **Estimate the treatment mechanism $g_0(A|W) = P_0(A|W)$, which is the conditional probability of having Dinosaur experience, given baseline covariates.** You are provided with background knowledge that the true conditional probability might depend on the following variables:

$$\left\{A,\ W1,\ W2,\ W3,\ W4,\ cos(\pi W2),\ sin(\pi W3),\ W4^2\right\}$$

(a) Create transformed variables `cosW2`, `sinW3` and `W4sq`.

```
> cosW2<- cos(pi*ObsData$W2)
```

(b) Create data frame `W` of baseline covariates and the transformed variables.

(c) Estimate $g_0(A|W)$ by running `SuperLearner`. Call this object `gHatSL`. Since we are estimating the treatment mechanism, specify `Y=ObsData$A` and the predictors as `X=W`. Use the same library.

```
> gHatSL<- SuperLearner(Y=ObsData$A, X=W, SL.library=SL.library, family="binomial")
```

(d) The predicted probability of being experienced, given the subject's baseline characteristics $g_n(A = 1|W)$, can be accessed with `gHatSL$SL.predict`

i. Assign the predicted probability of being experienced $g_n(A=1|W)$ to `gHat1W`:
```
> gHat1W<- gHatSL$SL.predict
```
ii. Assign the predicted probability of not being experienced $g_n(A=0|W)$ to `gHat0W`.
iii. Look at the distribution of propensity scores $g_n(A=1|W)$ and $g_n(A=0|W)$.
iv. Generate the predicted probability of the observed exposure, given the subject's baseline characteristics $g_n(a|w)$.
- *Hint:* Create empty vector `gHatAW`. Among subjects with $A=1$, assign the predicted probabilities $g_n(A=1|W)$. Among subjects with $A=0$, assign the predicted probabilities $g_n(A=0|W)$.

(e) Evaluate the IPTW estimator by taking the empirical mean of the weighted observations:

$$\hat{\Psi}_{IPTW}(P_n) = \frac{1}{n}\sum_{i=1}^{n}\frac{\mathbb{I}(A_i=1)}{g_n(A_i|W_i)}Y_i - \frac{1}{n}\sum_{i=1}^{n}\frac{\mathbb{I}(A_i=0)}{g_n(A_i|W_i)}Y_i$$

- The first term is the empirical mean of the outcomes, where observations with $A_i=1$ (`as.numeric(ObsData$A==1)`) are weighted as the inverse of the predicted probability of the observed exposure, given the baseline covariates $1/g_n(A_i|W_i)$ and where observations with $A_i \neq 1$ are weighted 0.
- The second term is the empirical mean of the outcomes, where observations with $A_i=0$ (`as.numeric(ObsData$A==0)`) are weighted as the inverse of the predicted probability of the observed exposure, given the baseline covariates $1/g_n(A_i|W_i)$ and where observations with $A_i \neq 0$ are weighted 0.
- As before, this is not part of the TMLE algorithm, but implemented for comparison.

3. **Use these estimates to create the clever covariate:**

$$H_n(A,W) = \left(\frac{\mathbb{I}(A=1)}{g_n(A=1|W)} - \frac{\mathbb{I}(A=0)}{g_n(A=0|W)}\right)$$

(a) Calculate `H.AW` for each subject:
```
> H.AW<- as.numeric(ObsData$A==1)/gAW - as.numeric(ObsData$A==0)/gAW
```
For subjects with $A=1$, the clever covariate is 1 over the predicted probability of being experienced, given the baseline covariates. Among subjects with $A=0$, the clever covariate is -1 over the predicted probability of not being experienced, given the baseline covariates $g_n(A=0|W)$.

(b) Also evaluate the clever covariate at $A=1$ and $A=0$ for all subjects. Call the resulting values `H.1W` and `H.0W`, respectively.

4. **Update the initial estimates.**

(a) Run a logistic regression of the outcome Y on the clever covariate $H_n(A,W)$, using the logistic of the initial estimate as offset and suppressing the intercept.
```
> logitUpdate<- glm(ObsData$Y ~ -1 +offset(qlogis(QbarAW)) + H.AW, family='binomial')
```
- We suppress the intercept by including -1 on the right hand side.
- In R, logistic function is given by `qlogis(x)`.
- As always, including `family='binomial'` runs logistic regression.

(b) Let `eps` denote the resulting maximum likelihood estimate of the coefficient on the clever covariate `H.AW`.
```
> eps<- logitUpdate$coef
```

(c) Update the initial estimate of $\bar{Q}_n(A,W)$ according to the fluctuation model:

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

Create `QbarAW.star` by taking the inverse logit (i.e. the expit) of the offset (`logit(QbarAW)`) plus the coefficient ϵ_n times the clever covariate $H_n(A,W)$:

```
> QbarAW.star<- plogis(qlogis(QbarAW)+ eps*H.AW)
```

(d) Update the initial estimates of $\bar{Q}_n(1, W)$ and $\bar{Q}_n(0, W)$:

$$logit\big[\bar{Q}_n^*(1, W_i)\big] = logit\big[\bar{Q}_n(1, W_i)\big] + \epsilon_n H_n(1, W_i)$$
$$logit\big[\bar{Q}_n^*(0, W_i)\big] = logit\big[\bar{Q}_n(0, W_i)\big] + \epsilon_n H_n(0, W_i)$$

(e) *Optional*: Try updating again. What is updated ϵ_n?

5. **Substitute the updated fits into the target parameter mapping:**

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

5 The basics of the `tmle` package

1. **If you have not already, download and install the `tmle` package.**
2. **Load the package with `library("tmle")`.**
3. **Read the help file: `?tmle`:**
   - The basic input to the function `tmle` is the outcome `Y`, the intervention `A` and the baseline covariates `W` (need to be in a matrix or data frame).
   - The user can also specify mediating variables with `Z`, missing values with `Delta` and repeated observations with `id`.
   - The initial estimates of $\bar{Q}_0(A, W)$ can be supplied by the user in a nx2 matrix `Q`, can be estimated according to a user-specified regression formula `Qform`, or estimated with SuperLearner with library `Q.SL.library`.
   - The initial estimates of $g_0(A = 1|W)$ can be supplied by the user in a nx1 vector `g1W`, can be estimated according to a user-specified regression formula `gform`, or estimated with SuperLearner with library `g.SL.library`.
4. **Call the `tmle` function using the SuperLearner initial estimates for the conditional mean outcome $\bar{Q}_0(A, W)$ and for the treatment mechanism $g_0(A|W)$. Also, specify the family as binomial.**

   ```
   > tmle(Y=ObsData$Y, A=ObsData$A, W=W, Q=cbind(QbarOW, Qbar1W), g1W=gHat1W, family="binomial")
   ```

   Recall that the data frame `W` of includes baseline covariates and the transformed variables.
5. **Use the `summary` and `names` functions to explore the output.**

6 Use the `tmle` package to explore performance under model misspecification

1. **Implement `tmle` with the correctly specified model for $\bar{Q}_0(A, W)$ and for $g_0(A|W)$. Specify the regression formulas in `Qform` and `gform`.** The true conditional probability of surviving, given the exposure and baseline covariates, is described by the following parametric model

$$logit[\bar{Q}_0(A, W)] = \beta_0 + \beta_1 A + \beta_2 cos(\pi W2) + \beta_3 sin(\pi W3) + \beta_4 W4^2$$

The true conditional probability of the exposure (having prior Dinosaur experience), given the baseline covariates, is described by the following parametric model

$$logit[g_0(A = 1|W)] = \beta_0 + \beta_1 cos(\pi W2) + \beta_2 sin(\pi W3) + \beta_4 W1^* W4$$

```
> tmle.Qcorr.gcorr<- tmle(Y=ObsData$Y, A=ObsData$A, W=W,
+   Qform=Y~ A+cosW2+sinW3+W4sq, gform=A~cosW2+sinW3+W1:W4, family="binomial")
```

- Using `W1:W4` ensures that only the interaction between $W1$ and $W4$ is included in the model for $g_0(A|W)$.

2. **Implement `tmle` with a misspecified model for $\bar{Q}_0(A,W)$ and a correctly specified model for $g_0(A|W)$.**
3. **Implement `tmle` with a correctly specified model for $\bar{Q}_0(A,W)$ and a misspecified model for $g_0(A|W)$.**
4. **Implement `tmle` with a misspecified model for $\bar{Q}_0(A,W)$ and a misspecified model for $g_0(A|W)$.**
5. **Implement `tmle` using SuperLearner with the default library for initial estimates of $\bar{Q}_0(A,W)$ and $g_0(A|W)$.**
6. **Compare the resulting point estimates for $\Psi(P_0)$ for this sample of $n = 5,000$ subjects.**

Note: There is a new TMLE package for point treatment as well as longitudinal problems: `ltmle`.