R Lab 4 - IPTW & the Positivity Assumption

Introduction to Causal Inference

Goals:
1. Implement IPTW for statistical parameter equal to the average treatment effect under the necessary causal assumptions.
2. Implement IPTW for parameters of a MSM.
3. Understand how the IPTW estimator is affected by "near" positivity violations and weight stabilization.

Next lab:
We will implement targeted maximum likelihood estimation (TMLE).

1 Background Story

Given our previous success at predicting whether or not a pirate ship will find buried treasure, the Queen has hired us to estimate the causal effect of scurvy on mortality among pirates. The data available on n pirates are

- $W1$: possession of at least one awesome pirate characteristic (e.g. peg leg, eye patch, beard, parrot). This is a binary covariate measured when leaving port (yes:1, no:0).
- $W2$: a summary measure of route danger, including voyage length, hurricane season, travel through enemy waters, Bermuda triangle, etc. This is a categorical variable ranging from 0 as least dangerous to 3 as most dangerous.
- A: whether the pirate suffered from scurvy during the voyage. This is a binary exposure (yes:1, no:0).
- Y: pirate's mortality status. This is a binary outcome with 1 for deceased and 0 for alive.

1.1 Causal Road Map Rundown

Please note: We are doing a very fast review here. In practice, each step of the road map requires very careful thinking.

1. **Specify the Question:**
   What is the causal effect of scurvy on mortality among pirates?

2. **Specify the causal model:**
   - Endogenous nodes: $X = (W1, W2, A, Y)$, where $W1$ is an indicator for having an "awesome" pirate characteristic, $W2$ is a summary measure of route danger, A is scurvy status and Y is mortality.
   - Exogenous nodes: $U = (U_{W1}, U_{W2}, U_A, U_Y) \sim P_U$. We make no assumptions about the distribution P_U.
   - Structural equations F:

$$\begin{aligned} W1 &= f_{W1}(U_{W1}) \\ W2 &= f_{W2}(W1, U_{W2}) \\ A &= f_A(W1, W2, U_A) \\ Y &= f_Y(W1, W2, A, U_Y) \end{aligned}$$

   There are no exclusion restrictions or assumptions about functional form.

Image 1: https://www.flickr.com/photos/talklikeapirateday/3933458622/in/set-990505
Image 2: http://www.huffingtonpost.com/2013/09/24/sir-stuffington-one-eyed-kitten_n_3982907.html

3. **Specify the causal parameter of interest:**
   We are interested in the causal risk of death due to scurvy (i.e. the average treatment effect):
   $$\Psi^F(P_{U,X}) = E_{U,X}(Y_1) - E_{U,X}(Y_0) = P_{U,X}(Y_1 = 1) - P_{U,X}(Y_0 = 1)$$
   where Y_a denotes the counterfactual outcome (mortality), if possibly contrary to fact, the pirate had scurvy status $A = a$.

4. **Specify the link between the SCM and the observed data:**
   The observed data were generated by sampling n independent times from a data generating system compatible with the structural causal model $\mathcal{M}^F$. This yield n i.i.d. copies of random variable $O = (W1, W2, A, Y) \sim P_0$. 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. A sufficient, but not minimal, identifiability assumption is that all of the exogenous errors are independent. Other possibilities include $U_A \perp\!\!\!\perp U_Y$ and (i) $U_A \perp\!\!\!\perp U_{W1}$, $U_A \perp\!\!\!\perp U_{W2}$ or (ii) $U_Y \perp\!\!\!\perp U_{W1}$, $U_Y \perp\!\!\!\perp U_{W2}$. We use $\mathcal{M}^{F*}$ to denote the original SCM augmented by the additional causal assumptions needed for identifiability. We introduce this "working" SCM to keep our real knowledge separate from our wished for identifiability assumptions. Under $\mathcal{M}^{F*}$, the backdoor criteria holds conditionally on $W = (W1, W2)$.

   For identifiability, 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 words, we need that each possible exposure level is represented in every covariate strata. This condition on data support ensures that our statistical estimand is well-defined. Specifically, we need a non-zero probability of the conditioning set in the G-computation formula. In our example, there must be a positive probability of having scurvy or not, within strata of "awesome" pirate status and route danger. Here, we are using $W = (W1, W2)$ to denote the set of covariates that satisfy the backdoor criteria under the working SCM $\mathcal{M}^{F*}$.

6. **Specify the target parameter of the observed data distribution:**
   Under the working SCM $\mathcal{M}^{F*}$, the average treatment effect $\Psi^F(P_{U,X})$ is identified using the G-Computation formula:
   $$\begin{aligned}\Psi(P_0) &= E_0\big[E_0(Y|A = 1, W) - E_0(Y|A = 0W)\big] \\ &= \sum_w \big[E_0(Y|A = 1, W = w) - E_0(Y|A = 0, W = w)\big] P_0(W = w)\end{aligned}$$
   where $W = (W1, W2)$. This is our statistical estimand.

7. **Estimate the chosen parameter of the observed data distribution:**
   We have discussed two estimators of the statistical parameter. They rely on estimating different parts of the observed data distribution P_0:

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

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

   where P_n is the empirical distribution and $\bar{Q}_n(A, W)$ is the estimate of the conditional mean outcome given the exposure (scurvy) and baseline covariates $E_0(Y|A, W)$.

   (b) **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 $g_n(A_i|W_i)$ is an estimate of the conditional probability of scurvy given the baseline characteristics $P_0(A|W)$. This conditional distribution is often referred to as the exposure or treatment mechanism. IPTW is the focus of today's lab.

   (c) TMLE: coming soon :)

8. **Inference and interpret results:** Coming soon.

2 Import and explore data set `RLab4.IPTW.csv`.

1. Use the `read.csv` function to import the data set and assign it to data frame `ObsData`.
2. Use the `nrow` function to count the number of pirates in the data set. Assign this number as `n`.
3. Use the `names`, `tail` and `summary` functions to explore the data.
4. With the `table` function, the number of pirates in each covariate strata without scurvy $A = 0$ and the number of pirates in each covariate strata with scurvy $A = 1$. *Note: these tables are simply counting the number of observations within each strata of* $(W1, W2, A)$ *in a single sample of size* n*; we are not formally evaluating the positivity assumption, which is a statistical assumption on the true data generating process* P_0.

```
> table(ObsData$W1,ObsData$W2, ObsData$A)
```

3 Implement the IPTW for the G-computation estimand

1. Estimate the treatment mechanism $P_0(A|W) = g_0(A|W)$, which is the conditional probability of scurvy, given the pirate's characteristics. Use the following *a priori*-specified parametric regression model:

   $$g_0(A = 1|W) = expit\left[\beta_0 + \beta_1 W1 + \beta_2 W2\right]$$

   Hint: Run `glm` with specifications `family='binomial'` for logistic regression and `data=ObsData`.
   In practice, we would generally use a data-adaptive estimator, such as SuperLearner. To save time during lab, we are using the correctly specified parametric regression.

2. Predict each pirate's probability of his observed exposure (scurvy status), given his covariates: $g_n(A_i|W_i)$...

   (a) Obtain the predicted probability of having scurvy, given the baseline covariates `pred.g1W`.
   Hint: Apply the `predict` function to the above fitted logistic regression function with `type='response'`.

(b) Obtain the predicted probability of not having scurvy, given the baseline covariates `pred.g0W`:

$$g_n(0|W) = 1 - g_n(1|W)$$

(c) Create an empty vector `gAW` of length n.

(d) Among pirates with scurvy, assign the appropriate predicted probability:

```
> gAW[ObsData$A==1] <- pred.g1W[ObsData$A==1]
```

(e) Among pirates without scurvy, assign the appropriate predicted probability:

```
> gAW[ObsData$A==0] <- pred.g0W[ObsData$A==0]
```

(f) Use the `summary` function to examine the distribution of the predicted probabilities. Any cause for concern?

3. Create a vector `wt` as the inverse of the predicted probability. Use the `summary` function to examine the distribution of the weights.

4. Evaluate the IPTW estimand by taking the empirical mean of the weighted outcomes:

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

Hint: We can code indicator functions with `as.numeric` applied to a logical statement.

```
> #indicator that A_i=1 can be coded as
> as.numeric(ObsData$A==1)
```

5. Comment on the results.

6. Arbitrarily truncate the weights at 10 and evaluate the IPTW estimand.
Hint: The following code copies the weight vector (`wt`) into a new vector (`wt.trunc`) and truncates the weights at 10.

```
> wt.trunc<- wt
> wt.trunc[wt.trunc > 10]<- 10
```

7. Implement the stabilized IPTW estimator (a.k.a. the modified Horvitz-Thompson estimator):

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

Dividing by the mean of the weights ensures that the IPTW estimator is bounded. This formula may look intimidating. Just divide empirical mean of the weighted outcomes by the empirical mean of the weights.

4 IPTW & Marginal Structural Models

Recall: Marginal structural models (MSMs) are just another way to define the target parameter. Specifically, they provide a summary measure of how the expected counterfactual outcome Y_a changes as a function of treatment A and possibly effect modifiers of interest, denoted V.

Suppose the Queen is interested in how the expected counterfactual mortality Y_a varies as function of scurvy A and route danger $V = W2$. Specifically, she believes that the effect of scurvy on death is modified by the route danger. Consider the following MSM to summarize how the average counterfactual outcome mortality Y_a varies of as a function scurvy A and route danger V:

$$E_{U,X}(Y_a|V) = m(a, V|\beta) = expit\big[\beta_0 + \beta_1 a + \beta_2 V + \beta_3 a^* V\big]$$

For simplicity, we are treating this MSM as the truth. In class, we will cover estimation of parameters of *working* MSMs using IPTW with and without stabilized weights. This is step 1 of the causal road map: specifying the question. See `R` lab 1 for further discussion and worked examples.

4.1 Implement IPTW for the MSM parameter without stabilized weights:

1. Estimate the treatment mechanism $P_0(A|W) = g_0(A|W)$, which is the conditional probability of scurvy, given the pirate's characteristics.
   Hint: We already did this! Skip to next step.

2. Predict each pirate's probability of his observed exposure (scurvy status), given his covariates `gAW`.
   Hint: We already did this! Skip to next step.

3. Create the weight vector `wt` as the inverse of the predicted probabilities:

$$wt_i = \frac{g^*(A_i)}{g_n(A_i|W_i)}, \text{ where } g^*(A_i) = 1$$

   Hint: We already did this! Skip to the next step.

4. Estimate the β coefficients by regressing the outcome Y on the exposure A and effect modifier $V = W2$ according to the MSM.
   Hint: Use `glm` to run logistic regression. Specify the `weights`, the `family` and the `data`.

5. Interpret the results.

4.2 Weight stabilization in IPTW for an MSM parameter

For MSMs with effect modifiers $m(a, V|\beta)$, the numerator of the weights can be any function of the exposure A and the effect modifier V. A common choice is the conditional probability of the observed exposure, given the effect modifier:

$$st.wt_i = \frac{g_n^*(A_i|V_i)}{g_n(A_i|W_i)}, \text{ where } g_n^*(A_i|V_i) = P_n(A_i|V_i)$$

This can help reduce variability in the weights. If the covariates in W no longer predict the exposure (whether a pirate has scurvy) after controlling for the effect modifier (route's danger), then the weights will be 1. All the control for confounding is taken care by adjustment for $V = W2$ in the MSM. In less extreme cases, this choice of numerator can still increase efficiency. Pirates will get weighted based on the conditional probability of their observed exposure given the effect modifier $g_n(a|V)$ relative to the conditional probability of their observed exposure given all the baseline covariates $g_n(a|W)$.

The positivity assumption for a parameter defined with MSM $m(a, V|\beta)$ is also weakened:

$$sup_{a \in \mathcal{A}} \frac{g^*(a, V)}{g_0(a|w)} < \infty \text{ for all } w \text{ for which } P_0(W = w) > 0$$

For any (a, V) combinations that occur with a non-zero probability, we need that the conditional probability of that exposure a given the baseline covariates to also be non-zero. (We will cover this in more detail in class.)

4.3 Implement IPTW for a MSM parameter with stabilized weights:

1. Estimate the treatment mechanism $P_0(A|W) = g_0(A|W)$. *Hint: We already did this! Skip to next step.*

2. Predict each pirate's probability of his observed exposure (scurvy status), given his covariates `gAW`. *Hint: We already did this! Skip to next step.*

3. Create the stabilized weights `wt.MSM`:

$$st.wt_i = \frac{g_n^*(A_i|V_i)}{g_n(A_i|W_i)}, \text{ where } g_n^*(A_i|V_i) = P_n(A_i|V_i)$$

(a) Estimate the conditional probability of the scurvy, given route danger $g_0(A = 1|V)$ with a saturated logistic regression:

```
> gAV.reg <- glm(A ~ as.factor(W2), family="binomial", data=ObsData)
```

We are using the `as.factor` function to create dummy variables for each level of route danger $W2$.

(b) Obtain the predicted probability of having scurvy, given the route danger: `pred.g1V`.
Hint: Apply the `predict` function to the above fitted logistic regression function with `type='response'`.

(c) Calculate the predicted probability of not having scurvy, given the route danger: `pred.g0V`:

$$g_n(0|V) = 1 - g_n(1|V)$$

(d) Create an empty vector `gAV` of length n.

(e) Among pirates with scurvy, assign the appropriate predicted probability:

```
> gAV[ObsData$A==1] <- pred.g1V[ObsData$A==1]
```

(f) Among pirates without scurvy, assign the appropriate predicted probability:

```
> gAV[ObsData$A==0] <- pred.g0V[ObsData$A==0]
```

(g) Create the stabilized weights:

```
> wt.MSM<- gAV/gAW
```

(h) Look at the distribution of the stabilized weights.

4. Estimate the parameters of the MSM by regressing the observed outcome Y on the exposure A and effect modifier V according to the MSM. Specify the `weights`, the `family` and the `data`.

5. Comment on the resulting estimates.