R Lab 3 - IPTW & the Positivity Assumption

Laura B. Balzer

Biostat 683- Intro. to Causal Inference

Goals:
1. Implement IPTW for a binary exposure.
2. Understand how the IPTW estimator is affected by "near" positivity violations and weight stabilization.
3. Extend IPTW to control for missingness on the outcome.

Next lab:
Code discrete Super Learner to select the estimator with the lowest cross-validated risk. Use `R SuperLearner` package to build the best convex combination of candidate algorithms and to evaluate the performance of Super Learner.

1 Background Story

You have been hired by the Queen to estimate the causal effect of scurvy on mortality among pirates. The data available on n pirates are

- $W1$: possession of at least one quintessential 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 upon returing to port. This is a binary outcome with 1 for deceased and 0 for alive.

1.1 Causal Roadmap 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 status.
   - Background variables: $U = (U_{W1}, U_{W2}, U_A, U_Y) \sim \mathbb{P}_U$. We make no assumptions about the distribution $\mathbb{P}_U$.

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

- 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.

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):

$$\Psi^*(\mathbb{P}^*) = \mathbb{E}^*(Y_1) - \mathbb{E}^*(Y_0) = \mathbb{P}^*(Y_1 = 1) - \mathbb{P}^*(Y_0 = 1)$$

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 $\mathcal{M}^*$. This yields n i.i.d. copies of random variable $O = (W1, W2, A, Y) \sim \mathbb{P}_0$. The statistical model $\mathcal{M}$ for the set of allowed distributions of the observed data is non-parametric.

5. **Assess identifiability:**
The target causal parameter is not identified from the observed data distribution. There are several unblockable backdoor paths from the outcome Y into the exposure A. For identifiability to hold, we would need $U_A \perp\!\!\!\perp U_Y$ and (i) $U_A \perp\!\!\!\perp U_{W1}$, $U_A \perp\!\!\!\perp U_{W2}$ or (ii) $U_Y \perp\!\!\!\perp U_{W1}$, $U_Y \perp\!\!\!\perp U_{W2}$.

6. **Specify the target parameter of the observed data distribution:**
Despite lack of identifiability, we can still "commit" to an interesting statistical estimand inspired by our scientific/causal question. Let $W = (W1, W2)$ denote our adjustment set; then our statistical estimand is

$$\Psi(\mathbb{P}_0) = \mathbb{E}_0\big[\mathbb{E}_0(Y|A = 1, W)\big] - \mathbb{E}_0\big[\mathbb{E}_0(Y|A = 0, W)\big]$$

where the outer expectation is over the distribution of adjustment variables W. This can equivalently be expressed as the IPW estimand:

$$\Psi(\mathbb{P}_0) = \mathbb{E}_0\left[\frac{\mathbb{I}(A = 1)}{\mathbb{P}_0(A = 1|W)} Y\right] - \mathbb{E}_0\left[\frac{\mathbb{I}(A = 0)}{\mathbb{P}_0(A = 0|W)} Y\right]$$

where the second equality holds due to the linearity of expectations.

For identifiability, we also need the positivity assumption to hold:

$$min_{a\in\mathcal{A}}\ \mathbb{P}_0(A=a|W=w)>0$$

for all w for which $\mathbb{P}_0(W=w)>0$. 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. In our example, there must be a positive probability of having scurvy or not, within strata of "awesome" pirate status and route danger.

We have not changed our statistical model $\mathcal{M}$, which remains non-parametric.

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 $\mathbb{P}_0$:

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

   $$\hat{\Psi}_{SS}(\mathbb{P}_n)=\frac{1}{n}\sum_{i=1}^{n}\hat{\mathbb{E}}(Y|A=1,W_i)-\frac{1}{n}\sum_{i=1}^{n}\hat{\mathbb{E}}(Y|A=0,W_i)$$

   where $\mathbb{P}_n$ is the empirical distribution and $\hat{\mathbb{E}}(Y|A,W)$ is the estimate of the conditional mean outcome given the exposure (scurvy) and adjustment variables $\mathbb{E}_0(Y|A,W)$.

   (b) **Inverse probability weighted estimator (IPTW)**:

   $$\begin{aligned}\hat{\Psi}_{IPTW}(\mathbb{P}_n)&=\frac{1}{n}\sum_{i=1}^{n}\left(\frac{\mathbb{I}(A_i=1)}{\hat{\mathbb{P}}(A=1|W_i)}-\frac{\mathbb{I}(A_i=0)}{\hat{\mathbb{P}}(A=0|W_i)}\right)Y_i\\&=\frac{1}{n}\sum_{i=1}^{n}\left(\frac{\mathbb{I}(A_i=1)}{\hat{\mathbb{P}}(A=1|W_i)}\right)Y_i-\frac{1}{n}\sum_{i=1}^{n}\left(\frac{\mathbb{I}(A_i=0)}{\hat{\mathbb{P}}(A=0|W_i)}\right)Y_i\end{aligned}$$

   where $\hat{\mathbb{P}}(A=1|W_i)$ is an estimate of the conditional probability of scurvy, given pirate i's baseline characteristics $\mathbb{P}_0(A=1|W_i)$. This conditional probability is often referred to as the "propensity score". 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 `RLab3.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.**

3. **Use the `names`, `tail` and `summary` functions to explore the data.**

4. **With the `table` function, check the number of pirates in each covariate strata without scurvy $A=0$ and the number of pirates in each adjustment 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 structural positivity assumption, which is a statistical assumption on the true data generating process $\mathbb{P}_0$.*

```
> table(ObsData[,c('W1', 'W2', 'A')])
```

3 Implement IPTW for a binary exposure

1. **Estimate the propensity score $\mathbb{P}_0(A = 1|W)$, which is the conditional probability of scurvy, given the pirate's characteristics. Use the following *a priori*-specified parametric regression model:**

$$\mathbb{P}_0(A = 1|W) = logit^{-1}\big[\beta_0 + \beta_1 W1 + \beta_2 W2\big]$$

Hint: Run `glm` with specifications `family='binomial'` for logistic regression and `data=ObsData`.
In practice, we would generally use a machine learning algorithm, such as Super Learner (coming next).

2. **Predict each pirate's probability of having and not having scurvy, given their covariates: $\hat{\mathbb{P}}(A = 1|W_i)$ and $\hat{\mathbb{P}}(A = 0|W_i)$.**

   (a) Obtain the predicted probability of having scurvy, given the baseline covariates `prob.1W`.
   Hint: Apply the `predict` function to the above fitted logistic regression function with `type='response'`.

   (b) Also obtain the predicted probability of not having scurvy, given the baseline covariates `prob.0W`:

$$\hat{\mathbb{P}}(A = 0|W_i) = 1 - \hat{\mathbb{P}}(A = 1|W_i)$$

   (c) Use the `summary` function to examine the distribution of the predicted probabilities `prob.1W` and `prob.0W`. Any cause for concern?

3. **Create the weights for each pirate as an indicator of receiving the exposure of interest, divided by their predicted probability of receiving that exposure**

   (a) Create a vector `wt1` as an indicator of having scurvy, divided by the estimated probability of having scurvy given the adjustment set: $\mathbb{I}(A_i = 1)/\hat{\mathbb{P}}(A = 1|W_i)$.

```
> wt1 <- as.numeric(ObsData$A==1)/prob.1W
```

   where we are coding the indicator function with `as.numeric` applied to a logical statement.

   (b) Create a vector `wt0` as an indicator of noth having scurvy, divided by the estimated probability of not having scurvy given the adjustment set: $\mathbb{I}(A_i = 0)/\hat{\mathbb{P}}(A = 0|W_i)$.

```
> wt0 <- as.numeric(ObsData$A==0)/prob.0W
```

   (c) Comment on the distribution of the weights.

4. **Evaluate the IPTW estimand by taking the difference of the empirical means of the weighted outcomes:**

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

```
> mean(wt1*ObsData$Y) - mean(wt0*ObsData$Y)
>
```

This is equivalent to

```
> mean( (wt1-wt0)*ObsData$Y)
>
```

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 (`wt1`) into a new vector (`wt1.trunc`) and truncates the weights at 10.

```
> wt1.trunc<- wt1
> wt1.trunc[wt1.trunc > 10]<- 10
```

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

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

Dividing by the mean of the weights ensures that the IPTW estimator is bounded.

As shown in the Appendix, we were using the correctly specified regression to estimate the propensity score - and still saw a big between the true value of 28.7% and our estimates. To evaluate the performance of these 3 IPTW estimators, we could draw another independent sample of size n, implement the three estimators, and repeat 500+ times. Then we could evaluate the bias, variance, and mean squared error of these estimators for this data generating process. See* `R` *Lab2 and* `R` *homework 2.

4 Extensions to handle missingness

In the previous dataset, all pirates were magically followed to death or conclusion of their voyage - regardless of their characteristics or the route danger (e.g., Bermuda Triangle). More “realistically”, we have to deal with incomplete measurement of outcome. One approach to handle missingness is to modify our scientific question: “what would be the causal effect of scurvy on mortality among pirates under complete measurement of the outcome?”. As extra practice, you can work through the Causal Roadmap for this modified question. For the purposes of the `R` lab, we are going to skip to estimation.

In the following, let Δ be an indicator that the outcome was observed, and redefine the outcome Y equal to 1 if the pirate was observed/reported to have died and 0 otherwise (either did not die or had a missing outcome). Again let $W = (W1, W2)$ denote our adjustment set. We are now focused on estimating the following statistical estimand, corresponding to this causal question if the identifiability assumptions held:

$$\begin{aligned}\Psi(\mathbb{P}_0) &= \mathbb{E}_0\left[\frac{\mathbb{I}(A=1,\Delta=1)}{\mathbb{P}_0(A=1,\Delta=1 \mid W)}Y\right] - \mathbb{E}_0\left[\frac{\mathbb{I}(A=0,\Delta=1)}{\mathbb{P}_0(A=0,\Delta=1 \mid W)}Y\right] \\ &= \mathbb{E}_0\left[\frac{\mathbb{I}(A=1,\Delta=1)}{\mathbb{P}_0(A=1 \mid W)\mathbb{P}(\Delta=1 \mid A,W)}Y\right] - \mathbb{E}_0\left[\frac{\mathbb{I}(A=0,\Delta=1)}{\mathbb{P}_0(A=0 \mid W)\mathbb{P}_0(\Delta=1 \mid A,W)}Y\right]\end{aligned}$$

where in the second equality, we factored the denominator of the weights according to the assumed time-ordering: the exposure of scurvy happens before measurement/missingness on the outcome.

1. **Use the** `read.csv` **function to import the modified data** `RLab3.IPTW.missing.csv`**, and assign it to object** `ObsData`**.**
2. **Use the** `summary` **function to explore the dataset. How many pirates have their outcome measured?**
3. **Estimate the propensity score $\mathbb{P}_0(A=1|W)$, which is the conditional probability of scurvy, given the pirate's characteristics. Use the following *a priori*-specified parametric regression model:**

   $$\mathbb{P}_0(A=1|W) = logit^{-1}\left[\beta_0 + \beta_1 W1 + \beta_2 W2\right]$$

   Hint: We already did this. Skip to the next step.
4. **Predict each pirate's probability of having and not having scurvy, given their personal characteristics, and route danger: $\hat{\mathbb{P}}(A=1|W_i)$ and $\hat{\mathbb{P}}(A=0|W_i)$.**
   Hint: We already did this; the vectors of the corresponding probabilies are given by `prob.1W` *and* `prob.0W`*, respectively. Skip to the next step.*

5. **Estimate the probability of being measured, given the exposure, personal characteristics, and route danger: $\mathbb{P}_0(\Delta = 1|A, W)$. Use the following *a priori*-specified parametric regression model:**

$$\mathbb{P}_0(\Delta = 1|A, W1, W2) = logit^{-1}\big[\beta_0 + \beta_1 W1 + \beta_2 W2 + \beta_3 A\big]$$

6. **Predict each pirate's probability of being measured, given their observed past $\hat{\mathbb{P}}(\Delta = 1|A_i, W_i)$. Name the vector of the resulting probabilities `prob.delta`. Use the `summary` function to examine the distribution of `prob.delta`. Any cause for concern?**
   Hint: Apply the `predict` function to the above fitted logistic regression function with `type='response'`.

7. **Create the weights for each pirate as an indicator of receiving the exposure of interest *and being measured*, divided by their predicted probability of receiving that exposure and being measured.**

   (a) Create a vector `wt1` with numerator as an indicator of having scurvy and being measured, and with denominator as the estimated probability of having scurvy, given the adjustment set, times the estimated probability of being measured, given the observed past:

$$wt1_i = \frac{\mathbb{I}(A_i = 1, \Delta_i = 1)}{\hat{\mathbb{P}}(A = 1|W_i) \times \hat{\mathbb{P}}(\Delta = 1|A_i, W_i)}$$

```
> wt1 <- as.numeric(ObsData$A==1 & ObsData$Delta==1)/(prob.1W*prob.delta)
```

   (b) Create a vector `wt0` with numerator as an indicator of not having scurvy and being measured, and with denominator as the estimated probability of not having scurvy, given the adjustment set, times the estimated probability of being measured, given the observed past:

$$wt0_i = \frac{\mathbb{I}(A_i = 0, \Delta_i = 1)}{\hat{\mathbb{P}}(A = 0|W_i) \times \hat{\mathbb{P}}(\Delta = 1|A_i, W_i)}$$

```
> wt0 <- as.numeric(ObsData$A==0 & ObsData$Delta==1)/(prob.0W*prob.delta)
```

   (c) Comment on the distribution of the weights.

8. **Evaluate the IPTW estimand by taking the difference of the empirical means of the weighted outcomes:**

$$\hat{\Psi}_{IPTW}(\mathbb{P}_n) = \frac{1}{n}\sum_{i=1}^{n} \frac{\mathbb{I}(A_i = 1, \Delta_i = 1)}{\hat{\mathbb{P}}(A = 1|W_i)\hat{\mathbb{P}}(\Delta = 1|A_i, W_i)} Y_i - \frac{1}{n}\sum_{i=1}^{n} \frac{\mathbb{I}(A_i = 0, \Delta_i = 1)}{\hat{\mathbb{P}}(A = 0|W_i)\hat{\mathbb{P}}(\Delta = 1|A_i, W_i)} Y_i$$

9. **Arbitrarily truncate the weights at 10 and evaluate the IPTW estimand.**

10. **Implement the stabilized IPTW estimator (a.k.a., the modified Horvitz-Thompson estimator).**

11. **Comment on the results.**

Again, to evaluate the performance of these 3 IPTW estimators, we should draw another independent sample of size n, implement the three estimators, and repeat 500+ times. Then we could evaluate the bias, variance, and mean squared error of these estimators for this data generating process. See `R` Lab2 and `R` homework 2.