R Assignment 3 - IPTW

Laura B. Balzer

Biostat683 - Intro. to Causal Inference

Assigned: November 1, 2021
Write-ups due: Uploaded to your personal GoogleDrive folder by November 10, 2021 by 2:30pm. 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. Use of RMarkdown is strongly encouraged.

1 Background and Causal Roadmap

Dog People Live Longer. But Why? - *NPR*
"The studies, published in the journal *Circulation: Cardiovascular Quality and Outcomes*, suggest that dog ownership is linked to a 21% reduction in the risk of death - over the 12-year period studied - for people with heart disease. Those studies, along with a body of literature linking dogs to good health, all point toward one thing, says Dr. Dhruv Kazi... 'When you look at the big picture and look at all the evidence around dog ownership and cardiovascular health, it's pretty clear the signal is real and likely causal."'
https://www.npr.org/sections/health-shots/2019/10/26/773531999/dog-people-live-longer-but-why

Suppose our goal is understand the effect of dog ownership on subsequent mortality among older adults with cardiovascular disease. We have data on the the following variables:

- $W1$: Indicator of living in a rural area
- $W2$: Age in years
- $W3$: Centered measure of cardiovascular health at the study's start
- $W4$: Centered measure of socioeconomic status (SES)
- A: Indicator of having a dog at the study's start
- Y: Indicator of death within 12-years

Causal Roadmap Rundown

This is a very, very quick summary for review. Each step of the roadmap requires careful thought and consideration.

1. **Specify the Question:**
   What is the causal effect of having a dog on subsequent mortality among older adults with cardiovascular disease?

2. **Specify the causal model:**
   - Endogenous nodes: $X = (W1, W2, W3, W4A, Y)$
   - Background variables: $U = (U_{W1}, U_{W2}, U_{W3}, U_{W4}, U_A, U_Y) \sim \mathbb{P}_U$. We make no assumptions about the distribution $\mathbb{P}_U$.

- Structural equations F:

$$\begin{aligned}
W1 &= f_{W1}(U_{W1})\\
W2 &= f_{W2}(W1, U_{W2})\\
W4 &= f_{W4}(W1, W2, U_{W4})\\
W3 &= f_{W3}(W1, W2, W4, U_{W3})\\
A &= f_A(W1, W2, W3, W4, U_A)\\
Y &= f_Y(W1, W2, W3, W4, A, U_Y)
\end{aligned}$$

There are no exclusion restrictions or assumptions about functional form.

3. **Specify the causal parameter of interest:**
We are interested in the difference in the counterfactual risk of death if all older adults with cardiovascular disease had versus did not have a dog:

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

where Y_1 denotes the counterfactual outcome (mortality), if possibly contrary to fact, the older adult had a dog $A = 1$, and where Y_0 denotes the counterfactual outcome (mortality), if possibly contrary to fact, that same older adult did not have a dog $A = 0$.

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 yield n i.i.d. copies of random variable $O = (W1, W2, W3, W4, A, Y) \sim \mathbb{P}_0$. 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 (death) to the exposure (having a dog). For identifiability to hold, we would need the randomization assumption to hold:

$$Y_a \perp\!\!\!\perp A \mid \big(W1, W2, W3, W4\big)$$

In words, we need counterfactual survival to be independent from the observed exposure, given rurality, age, baseline cardiovascular health, and SES.

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 $W = (W1, W2, W3, W4)$ denote our adjustment set; then our statistical estimand is

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

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 regardless of rural/urban living, age, baseline cardiovascular health, and SES, there is a positive probability of having and not having a dog. This condition on data support ensures that our statistical estimand is well-defined.

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

2 Implement IPTW for a binary exposure

1. **Read-in and explore the data set `RAssign3.csv`.**

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

$$\mathbb{P}_0(A = 1|W) = logit^{-1}\big[\beta_0 + \beta_1 W1 + \beta_2 W2 + \beta_3 W3 + \beta_4 W4\big]$$

   In practice, we would generally use a machine learning algorithm, such as Super Learner (coming next).

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

4. **Use the `summary` function to examine the distribution of the predicted probabilities $\hat{\mathbb{P}}(A = 1|W_i)$ and $\hat{\mathbb{P}}(A = 0|W_i)$. Any cause for concern?**

5. **Create the weights, and comment on the distribution of the weights.**

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

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

7. **Arbitrarily truncate the weights at 10 and re-evaluate the IPTW estimand.**

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

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

9. **For comparison, also implement the unadjusted estimator.**

$$\begin{aligned}\hat{\Psi}_{unadj}(\mathbb{P}_n) &= \hat{\mathbb{E}}(Y|A = 1) - \hat{\mathbb{E}}(Y|A = 0)\\ &= \frac{1}{n}\sum_{i=1}^{n}\frac{\mathbb{I}(A_i = 1)}{\hat{\mathbb{P}}(A = 1)}Y_i \; - \; \frac{1}{n}\sum_{i=1}^{n}\frac{\mathbb{I}(A_i = 0)}{\hat{\mathbb{P}}(A = 0)}Y_i\end{aligned}$$

10. ***Bonus:* Implement a simple substituion estimator (a.k.a., parametric G-computaion) of $\Psi(\mathbb{P}_0)$ using the following parametric regression to estimate $\mathbb{E}_0(Y \mid A, W1, W2, W3, W4)$:**

$$\mathbb{E}(Y|A, W1, W2, W3, W4) = logit^{-1}\big[\beta_0 + \beta_1 W1 + \beta_2 W2 + \beta_3 W3 + \beta_4 W4 + \beta_5 A\big]$$

11. **Comment on the results.**

3 Extensions to handle missingness

In the following, let Δ be an indicator that the outcome (mortality) was observed, and redefine the outcome Y equal to 1 if the older adult was observed/reported to have died and 0 otherwise (either did not die or had a missing outcome). Again let $W = (W1, W2, W3, W4)$ denote our adjustment set. Now focus on the following statistical estimand, which controls for measured confounding by W as well as incomplete measurement of the outcome:

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

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. **Import and explore the modified data set `RAssign3.missing.csv`.**

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

$$\mathbb{P}_0(A=1|W) = logit^{-1}\big[\beta_0 + \beta_1 W1 + \beta_2 W2 + \beta_3 W3 + \beta_4 W4\big]$$

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

4. **Estimate the probability of being measured, given the exposure, rural/urban living, age, baseline cardiovascular health, and SES: $\mathbb{P}_0(\Delta=1|A,W)$. Use the following *a priori*-specified parametric regression model:**

$$\mathbb{P}_0(\Delta=1|A,W1,W2) = logit^{-1}\big[\beta_0 + \beta_1 W1 + \beta_2 W2 + \beta_3 W3 + \beta_4 W4 + \beta_5 A\big]$$

5. **Predict each participants's probability of being measured, given their observed past $\hat{\mathbb{P}}(\Delta = 1|A_i,W_i)$.**

6. **Create the weights - now accounting for confounding and incomplete measurement**

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

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

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

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

   (c) Comment on the distribution of the weights.

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

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

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

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

10. **For comparison, also implement the unadjusted estimator.**

$$\begin{aligned}\hat{\Psi}_{unadj}(\mathbb{P}_n) &= \hat{\mathbb{E}}(Y|A=1,\Delta=1) - \hat{\mathbb{E}}(Y|A=0,\Delta=1)\\ &= \frac{1}{n}\sum_{i=1}^{n}\frac{\mathbb{I}(A_i=1,\Delta_i=1)}{\hat{\mathbb{P}}(A=1,\Delta=1)}Y_i \; - \; \frac{1}{n}\sum_{i=1}^{n}\frac{\mathbb{I}(A_i=0,\Delta_i=1)}{\hat{\mathbb{P}}(A=0,\Delta=1)}Y_i\end{aligned}$$

11. ***Bonus:*** **Implement a simple substituion estimator (a.k.a., parametric G-computation) of $\Psi(\mathbb{P}_0)$ where in the first step the following parametric regression is used to estimate $\mathbb{E}_0(Y \mid A, W1, W2, W3, W4)$ - *among those who are measured*:**

$$\mathbb{E}(Y|A,\Delta=1,W1,W2,W3,W4) = logit^{-1}\big[\beta_0 + \beta_1 W1 + \beta_2 W2 + \beta_3 W3 + \beta_4 W4 + \beta_5 A\big]$$

```
> outcome.reg <- glm(Y ~ A + W1 +W2 +W3 +W4, data=ObsData[ObsData$Delta==1,],
+                    family='binomial')
```

12. **Comment on the results.**

4 Improving IPTW - Unrelated story to #DogsAndDAGs

This section uses the data generating distribution given in `Rassign3_modifiedIPTW.R`. In particular, the data generating distribution $\mathbb{P}_0$ is given by

$$\begin{aligned} W &\sim Bernoulli(.5)\\ A \mid W &\sim Bernoulli(0.2 + 0.6 \times W)\\ Y \mid A, W \;&=_d 1000 \;+ \mathbb{I}(\tilde{U} < logit^{-1}(W \times A)),\end{aligned}$$

where $\tilde{U} \sim Uniform(0,1)$ is independent of the other variables and where $=_d$ indicates "has the same distribution as". **Note that Y only takes on the values 1000 and 1001.**

Our goal is to estimate

$$\begin{aligned}\Psi(\mathbb{P}_0) &= \sum_w \mathbb{E}_0[Y|A=1,W=w]\mathbb{P}_0(W=w)\\ &= \mathbb{E}_0\left[\frac{A}{\mathbb{P}_0(A=1|W)}Y\right]\end{aligned}$$

Since we are only interested in the exposed level, we can replace $\mathbb{I}(A=1)$ with simply A in the numerator of the IPTW estimand.

The file `Rassign3_modifiedIPTW.R` also implements the IPTW estimator and modified Horvitz-Thompson estimator (i.e., stablized IPTW) of $\Psi(\mathbb{P}_0)$. In this problem we will assume that $\mathbb{P}_0(A=1|W)$ is known to the investigators (as in a randomized controlled trial without missingness). These estimators are then given by:

$$\begin{aligned}\hat{\Psi}_{IPTW}(\mathbb{P}_n) &= \frac{1}{n}\sum_{i=1}^{n}\frac{A_i}{\mathbb{P}_0(A=1|W_i)}Y_i\\ \hat{\Psi}_{HT}(\mathbb{P}_n) &= \frac{\sum_{i=1}^{n}\frac{A_i}{\mathbb{P}_0(A=1|W_i)}Y_i}{\sum_{i=1}^{n}\frac{A_i}{\mathbb{P}_0(A=1|W_i)}}\end{aligned}$$

In class we discussed how the modified Horvitz-Thompson estimator will often (although not always) yield finite sample improvements to the standard IPTW estimator. We also alluded to the fact that the modified

Horvitz-Thompson is asymptotically the same as the standard IPTW estimator; so the two will have similar behavior in large samples.

The code in `Rassign3_modifiedIPTW.R` also contains a space for `my.est`, an estimator that you will define and implement in this section. In particular, we will seek to modify the standard IPTW estimator in a different way to yield both finite sample and asymptotic improvements. The goal is for you to come up with (at least a precursor to) this estimator on your own. In the solution key, we will present the best possible modification to the IPTW estimator in terms of asymptotic performance. In class we will see that this estimator is asymptotically equivalent to the targeted maximum likelihood estimator (TMLE) for $\Psi(\mathbb{P}_0)$. Nonetheless, we expect the TMLE to perform better in finite samples for reasons that will be described in class.

Please complete Questions 1 through 7 listed below.

1. **Run the code given in `Rassign3_modifiedIPTW.R` and report how the standard IPTW and modified Horvitz-Thompson estimators perform in terms of bias, variance, and MSE over 2000 simulations each with sample size 1000. Which estimator would you use in practice?**
   `Note 1:` The estimator `my.est` will return `NA`, because you have not implemented it yet!
   `Note 2:` Both of these estimators are unbiased in finite samples when $\mathbb{P}_0(A = 1|W)$ is known; so any estimated bias is the result of only taking a finite number of Monte Carlo draws.

2. **Look at the IPTW column in the `est` matrix. What do you notice about the IPTW estimates across these 2000 Monte Carlo draws?**
   `Hint:` Recall the values that Y can take.

One way to address the problem that you described above is to use a modified Horvitz-Thompson estimator, which automatically respects the bounds of the statistical model and can lead to better finite sample performance. Nonetheless, there is another valid way to address this problem. Note that the IPTW estimate is an average of terms which are 0 (people with $A = 0$) and of terms which are larger than 1000 (people with $A = 1$). The calculations in the next set of questions will be useful for developing the intuition needed to create your own estimator.

3. **What is the variance of a random variable X with $\mathbb{P}(X = 0) = 1/2$ and $\mathbb{P}(X = 1) = 1/2$?**

4. **What is the variance of a random variable $X2$ with $Pr(X2 = 0) = 1/2$ and $Pr(X2 = 1000) = 1/2$?**

   `Hint:` $X2 = 1000 \times X$

5. **How are the above two calculations relevant to improving the IPTW estimator in this problem? We currently have an estimator that is an empirical mean of variables like those in Question 4. What transformation of the outcome Y would make your estimator behave more like an empirical mean of variables like those in Question 3?**

6. ***Graded leniently:* Write down an estimator $\hat{\Psi}_{my.est}$ which applies the ideas of the previous three questions into an estimator. There's no need to give the best possible estimator, but you should give an estimator that outperforms the IPTW estimator by a significant margin (i.e., does as or almost as well as the modified Horvitz-Thompson estimator in terms of bias/variance/MSE).**

7. ***Graded leniently:* Code your estimator and replace the NA on the line `my.est = NA` with the estimator you defined in the previous question. Report the bias/variance/MSE of your estimator over the 2000 Monte Carlo draws.**