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:**

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

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

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

11. **The true value is -27.3%. Comment on your results given your knowledge of the true value. If you completed the bonus, also include a discussion of the simple substitution estimator.**

Solution:

```
> # 1. Read in the data
> ObsData <- read.csv('RAssign3.csv')
> summary(ObsData)
```

```
       W1               W2              W3                W4                A
 Min.   :0.0000   Min.   :50.00   Min.   :-3.69000   Min.   :-3.49000   Min.   :0.0000
 1st Qu.:0.0000   1st Qu.:56.00   1st Qu.:-0.71250   1st Qu.:-0.66000   1st Qu.:0.0000
 Median :1.0000   Median :62.00   Median :-0.02000   Median : 0.03000   Median :0.0000
 Mean   :0.5028   Mean   :62.07   Mean   :-0.03393   Mean   : 0.01644   Mean   :0.3452
 3rd Qu.:1.0000   3rd Qu.:68.00   3rd Qu.: 0.63000   3rd Qu.: 0.68000   3rd Qu.:1.0000
 Max.   :1.0000   Max.   :75.00   Max.   : 3.78000   Max.   : 3.61000   Max.   :1.0000
       Y
 Min.   :0.000
 1st Qu.:0.000
 Median :0.000
 Mean   :0.292
 3rd Qu.:1.000
 Max.   :1.000

> dim(ObsData)

[1] 2500    6

> # 2.  Estimate the epxosure mechanism P(A|W) with W=(W1,W2,W3,W4)
> prob.AW.reg<- glm(A ~ W1 +W2 +W3 + W4, family="binomial", data=ObsData)

> # 3. # predicted probability of having a dog, given baseline characteristics
> prob.1W <- predict(prob.AW.reg, type= "response")
> # predicted probability of not having a dog, given baseline characteristics
> prob.0W <- 1 - prob.1W

> # 4. look at the distribution of predicted probabilities
> summary(prob.1W)

     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
0.0002953 0.0705011 0.2388129 0.3452000 0.5984808 0.9993384

> summary(prob.0W)

     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
0.0006616 0.4015192 0.7611871 0.6548000 0.9294989 0.9997047
```

IPTW is extremely sensitive to theoretical and practical positivity violations. From above summaries, we see that there are certain covariate combinations with little variability in the exposure (dog ownership).

```
> # 5. Create the weights
> wt1 <- as.numeric(ObsData$A==1)/prob.1W
> wt0 <- as.numeric(ObsData$A==0)/prob.0W
> summary(wt1)

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.000   0.000   0.000   1.046   1.188 177.719

> summary(wt0)

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.000   0.000   1.036   1.005   1.211  28.619
```

"Near" violations of the positivity assumptions often yield poor finite sample performance. Here, at least one older adult is being up-weighted by 178.

```
> # 6. Point estimate:
> iptw <- mean( wt1*ObsData$Y) -  mean( wt0*ObsData$Y)
> iptw

[1] -0.2416146

> # 7.  truncate  weights ARBITRARILY at 10
> # first let's see how many weights under the exposure are greater than 10
> sum(wt1>10)

[1] 21

> wt1.trunc<- wt1
> wt1.trunc[ wt1.trunc>10] <-10
> # same for weights under no exposure
> sum(wt0>10)

[1] 8

> wt0.trunc<- wt0
> wt0.trunc[ wt0.trunc>10] <-10
> # evaluate the IPTW estimand with the truncated weights
> iptw.tr <- mean( wt1.trunc*ObsData$Y) -  mean( wt0.trunc*ObsData$Y)
> iptw.tr

[1] -0.3195715

> # 8. Stabilized IPTW estimator - Modified Horvitz-Thompson estimator
> iptw.st <- sum( wt1*ObsData$Y)/sum( wt1) - sum( wt0*ObsData$Y)/sum( wt0)
> iptw.st

[1] -0.244704

> # 9. unadjusted
> unadj <- mean(ObsData[ObsData$A==1, 'Y']) - mean(ObsData[ObsData$A==0, 'Y'])
> unadj

[1] -0.4105452

> # same as
> mean( as.numeric(ObsData$A==1)/mean(ObsData$A)*ObsData$Y) -
+   mean( as.numeric(ObsData$A==0)/(1-mean(ObsData$A))*ObsData$Y)

[1] -0.4105452

> # 10 Bonus - Gcomp
> outcome.reg <- glm(Y ~ A + W1 +W2 +W3 +W4, data=ObsData, family='binomial')
> exp <- unexp <- ObsData
> exp$A <- 1
> unexp$A <- 0
> SS <- mean( predict(outcome.reg, newdata=exp, type='response') ) -
+   mean( predict(outcome.reg, newdata=unexp, type='response') )
> SS
```

```
[1] -0.2875429

> # 11. comparison
> round(data.frame(iptw, iptw.tr, iptw.st, unadj, SS)*100,1)

   iptw iptw.tr iptw.st unadj    SS
1 -24.2     -32   -24.5 -41.1 -28.8
```

11. The true value of the target parameter is -27.3%. For this single sample of $n = 2500$ older adults, the point estimates from standard IPTW, IPTW after truncating the weights, and stabilized IPTW were -24.2%, -32%, and -24.5%. For comparison, the unadjusted estimator, which does not control for measured confounding, is -41.1%, while the point estimate from the simple substitution estimator (a.k.a., parametric Gcomp.) was -28.8%.

We can interpret the statistical estimand $\Psi(\mathbb{P}_0)$ as the marginal difference in the mortality risk among older adults associated with having a dog, after controlling for rural/urban living, age, baseline cardiovascular health, and SES. If the identifiability assumptions (i.e., randomization and positivity) held, we could then interpret $\Psi(\mathbb{P}_0)$ in terms of the average treatment effect (a.k.a., the causal risk difference).

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:**

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

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:**

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

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 substitution 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*:**

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

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

12. **Comment on your results given your knowledge of the true value is -27.3%. If you completed the bonus, also include a discussion of the simple substitution estimator.**

Solution:

```
> # 1. read in data
> ObsData <- read.csv('RAssign3.missing.csv')
> summary(ObsData)
```

```
       W1               W2              W3                 W4                 A
 Min.   :0.0000   Min.   :50.00   Min.   :-3.69000   Min.   :-3.49000   Min.   :0.0000
 1st Qu.:0.0000   1st Qu.:56.00   1st Qu.:-0.71250   1st Qu.:-0.66000   1st Qu.:0.0000
 Median :1.0000   Median :62.00   Median :-0.02000   Median : 0.03000   Median :0.0000
 Mean   :0.5028   Mean   :62.07   Mean   :-0.03393   Mean   : 0.01644   Mean   :0.3452
 3rd Qu.:1.0000   3rd Qu.:68.00   3rd Qu.: 0.63000   3rd Qu.: 0.68000   3rd Qu.:1.0000
 Max.   :1.0000   Max.   :75.00   Max.   : 3.78000   Max.   : 3.61000   Max.   :1.0000
    Delta             Y
 Min.   :0.0000   Min.   :0.0000
 1st Qu.:1.0000   1st Qu.:0.0000
 Median :1.0000   Median :0.0000
 Mean   :0.8432   Mean   :0.2464
 3rd Qu.:1.0000   3rd Qu.:0.0000
 Max.   :1.0000   Max.   :1.0000

> # how many were measured
> sum(ObsData$Delta==1)

[1] 2108

> # 2. skip already done above
> prob.AW.reg

Call:  glm(formula = A ~ W1 + W2 + W3 + W4, family = "binomial", data = ObsData)

Coefficients:
(Intercept)           W1           W2           W3           W4
     0.9736      -0.7395      -0.0272       1.5555       1.5503

Degrees of Freedom: 2499 Total (i.e. Null);  2495 Residual
Null Deviance:            3222
Residual Deviance: 2017         AIC: 2027

> # 3. skip already done above
> summary(prob.1W); summary(prob.0W)

     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
0.0002953 0.0705011 0.2388129 0.3452000 0.5984808 0.9993384

     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
0.0006616 0.4015192 0.7611871 0.6548000 0.9294989 0.9997047

> # 4. estimate the probability of being measured, given the past
> prob.Delta.reg <- glm(Delta ~ W1 + W2 + W3 + W4 + A, family='binomial', data=ObsData)

> # 5. predicted probability of measurement, given the past
> prob.delta <- predict(prob.Delta.reg, type='response')
> summary(prob.delta)

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.03104 0.78941 0.92862 0.84320 0.97726 0.99987
```

```
> # 6. create the weights
> wt1 <- as.numeric(ObsData$A==1 & ObsData$Delta==1)/(prob.1W*prob.delta)
> wt0 <- as.numeric(ObsData$A==0 & ObsData$Delta==1)/(prob.0W*prob.delta)
> summary(wt1)

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.000   0.000   0.000   1.051   1.256 180.654

> summary(wt0)

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.000   0.000   1.027   1.049   1.273 102.405
```

"Near" violations of the positivity assumptions often yield poor finite sample performance. Here, at least one participant is being up-weighted by 180.6.

```
> # 7. Point estimate:
> iptw <- mean( wt1*ObsData$Y) - mean( wt0*ObsData$Y)

> # 8.  truncate  weights ARBITRARILY at 10
> sum(wt1>10)

[1] 23

> wt1.trunc<- wt1
> wt1.trunc[ wt1.trunc>10] <-10
> # same for weights under no exposure
> sum(wt0>10)

[1] 22

> wt0.trunc<- wt0
> wt0.trunc[ wt0.trunc>10] <-10
> # evaluate the IPTW estimand with the truncated weights
> iptw.tr <- mean( wt1.trunc*ObsData$Y) -  mean( wt0.trunc*ObsData$Y)

> # 9 Stabilized IPTW estimator - Modified Horvitz-Thompson estimator
> iptw.st <-  sum(wt1*ObsData$Y)/sum(wt1) - sum(wt0*ObsData$Y)/sum(wt0)

> # 10. Unadjusted
> exp.meas <- ObsData$A==1 & ObsData$Delta==1
> unexp.meas <- ObsData$A==0 & ObsData$Delta==1
> unadj <- mean( ObsData[exp.meas,'Y']) -  mean( ObsData[unexp.meas,'Y'])
> mean( as.numeric(exp.meas)/mean(exp.meas)*ObsData$Y) -
+   mean( as.numeric(unexp.meas)/mean(unexp.meas)*ObsData$Y)

[1] -0.4187345

> # 11 Bonus - Gcomp
> # Fit the outcome regression *among those with a measured outcome*
> # (i.e. with Delta=1)
> outcome.reg <- glm(Y ~ A + W1 +W2 +W3 +W4, data=ObsData[ObsData$Delta==1,],
```

```
+                    family='binomial')
> # Then implement the SS as usual
> # predict outcomes under A=1 and A=0 for all
> # then average and contrast
> exp <- unexp <- ObsData
> exp$A <- 1
> unexp$A <- 0
> SS <- mean( predict(outcome.reg, newdata=exp, type='response') ) -
+   mean( predict(outcome.reg, newdata=unexp, type='response') )

> # 12. comparison
> round(data.frame(iptw, iptw.tr, iptw.st, unadj, SS)*100,1)

   iptw iptw.tr iptw.st unadj    SS
1 -24.5   -31.8   -23.3 -41.9 -28.7
```

The true value of the target parameter is -27.3%. For this single sample of $n = 2500$ older adults among whom 2108 had their outcome observed, the point estimates from standard IPTW, IPTW after truncating the weights, and stabilized IPTW were -24.5%, -31.8%, and -23.3%. For comparison, the unadjusted estimator, which does not control for confounding or missingness, is -41.9%, while the point estimate from the simple substitution estimator (a.k.a., parametric Gcomp.) was -28.7%.

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.**

Solution:

1. The standard IPTW is orders of magnitude more variable that the modified HT etimator. This variance dominates the MSE. As noted in the instructions, both algorithms are unbiased, because the exposure mechanism is known. Overall, the modified/stabilized estimator has better performance and would be my choice.

   ```
   > source('Rassign3_iptwModified.R')
   ```

2. The IPTW estimator has a very high variance relative to the modified Horvitz-Thompson estimator. The minimum and maximum values of the estimates from IPTW are far from the true bounds of the outcome $Y \in \{1000, 1001\}$.

   ```
   > # 2. summary standard IPTW
   > summary(est[,1])

      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
     863.0   968.1  1000.6  1001.1  1034.4  1145.7
   ```

3. By properties of the binomial distribution, the variance of X is $p \times (1-p) = 0.5 \times (1-0.5) = 1/4$.

4. Now $X2$ is equal to 1000 times the random variable from the previous problem X. So the variance increases by a multiplicative factor of 1000^2. Thus, the variance is $Var(X2) = Var(1000 * X) = 1000^2 * Var(X) = 1000^2 \times 0.25 = 250,000$.

5. The IPTW estimator takes the average of terms which are

   - 0 when $A \neq 1$
   - $1/.8 \times Y \approx 1250$ when $A = 1$ and $W = 1$
   - $1/.2 \times Y \approx 5000$ when $A = 1$ and $W = 0$

   Because the marginal probability that $A = 1$ is equal to $\mathbb{P}_0(A = 1) = 1/2$, the IPTW estimator is an average of terms which are zero with probability $1/2$, and non-zero with probability $1/2$. Thus, the IPTW estimator is an empirical mean of random variables which will have even higher variance than the random variable $X2$ from the prevous problem, since 1250 and 5000 are larger than 1000. To reduce the variance, we will transform the weighted outcome to be bounded between 0 and 5. (The transformation is shown in the next step.)

6. We can use the estimator

   $$\hat{\Psi}_{my.est} = \frac{1}{n} \sum_{i=1}^{n} \frac{\mathbb{I}(A_i = 1)}{\mathbb{P}_0(A = 1 \mid W_i)} \big(Y_i - 1000\big) + 1000.$$

   That is, we subtract 1000 from our outcome, take the empirical mean of the weighted outcomes, and add 1000 to the final estimate. In this way we average over variable which take the values

   - 0 when $A = 0$ or when $(Y - 1000) = 0$
   - 1.25 when $A = 1$, $W = 1$, and $(Y - 1000) = 1$
   - 5 when $A = 1$, $W = 0$, and $(Y - 1000) = 1$

   Based on Questions 3 and 4, we expect that this average will have lower variance than the original IPTW estimator.

   In practice, it would be better to center the outcome about 0. Then our estimate would be behaving more like a random variable $X3$ where $Pr(X3 = -1/2) = Pr(X3 = 1/2) = 1/4$ and $Pr(X3 = 0) =$

1/2. To do this, we could subtract the sample mean of the outcome rather than the minimum value of the outcome (1000). This change will yield further variance gains.

Finally, we could also estimate the conditional mean outcome $\mathbb{E}_0(Y|A,W)$, then subtract the predicted outcome $\hat{\mathbb{E}}(Y|A_i,W_i)$ from each units's observed outcome Y_i. Under consistent estimation of $\mathbb{E}_0(Y|A,W)$ and $\mathbb{P}_0(A|W)$, this is will yield an asymptotically efficient estimator, i.e., will give the estimator with the lowest possible asymptotic variance among a large class of estimators. We will see in the next section that this double robust estimator also outperforms the modified Horvitz-Thompson estimator.

7. The estimators discussed above are coded below.

```
> set.seed(1)
> # The true value of the conditional mean outcome E_0[Y|A,W]
> true.meanY.AW <- function(A,W){
+   1000 + plogis(W*A)
+ }
> # The true value of propensity score Pr(A=1|W)
> true.prob.AW <- function(W){
+   0.2 + 0.6*W
+ }
> # A function which returns a data frame with n i.i.d. observations from P_0
> gen.data <- function(n){
+   # note this is a shortcut way of coding that skips generating the Us
+   # first and then generating the endogenous variables deterministically
+   W <- rbinom(n, 1, 1/2)
+   A <- rbinom(n, 1, true.prob.AW(W=W))
+   Y <- 1000 + rbinom(n, 1, true.meanY.AW(A=A,W=W) - 1000)
+   return(data.frame(W=W,A=A,Y=Y))
+ }
> # samples size
> n<- 1000
> # Number of Monte Carlo draws
> R <- 2000
> # Matrix of estimates from IPTW, modified Horvitz-Thompson, and my.est
> est <- matrix(NA,nrow=R,ncol=5)
> for(r in 1:R){
+   # Generate data with sample size
+   ObsData <- gen.data(n)
+   W <- ObsData$W
+   A <- ObsData$A
+   Y <- ObsData$Y
+   # True propensity score P_0(A=1|W)
+   pscore <- true.prob.AW(W=W)
+   # IPTW estimate
+   IPTW.est <- mean(A/pscore*Y)
+   # Modified Horvitz-Thompson estimate
+   HT.est <- mean(A/pscore*Y)/mean(A/pscore)
+   #
+   # Subtract minimum value of outcome (1000), run IPTW,
+   #   & add the min value back
+   my.est = mean(A/pscore*(Y-1000) + 1000)
+   #
+   # Subtract mean(Y), run IPTW, add mean(Y) back
+   my.est2 = mean(A/pscore*(Y-mean(Y)) + mean(Y))
```

```
+   # Double robust estimator.
+   # In practice would need to estimate E_0(Y|A,W)
+   my.est3 = mean(A/pscore*(Y-true.meanY.AW(A,W)) + true.meanY.AW(1,W))
+   est[r,] = c(IPTW.est, HT.est, my.est, my.est2, my.est3)
+ }
> #
> # Calculate the true value of sum_w E[Y|A=1,W=w) P(W=w)
> truth <- .5*true.meanY.AW(A=1, W=0) + .5*true.meanY.AW(A=1,W=1)
> # note: we know P_0(W=1) = 0.5
> truth

[1] 1000.616

> #
> # Calculate the estimated bias, variance, and MSE
> est.bias <- colMeans(est) - truth
> est.var <- apply(est,2,var)
> est.mse <- est.bias^2 + est.var
> #
> names(est.bias) <- names(est.var) <- names(est.mse) <-
+   c('IPTW','modifiedHT','Subtract1000','SubtractMean','DoubleRobust')
> #
> # The estimators have (estimated) bias:
> est.bias

         IPTW    modifiedHT  Subtract1000  SubtractMean  DoubleRobust
 4.990738e-01  2.180333e-04  3.238357e-04 -7.738690e-06 -3.338591e-05

> # The estimators have (estimated) variance:
> est.var

        IPTW   modifiedHT Subtract1000 SubtractMean DoubleRobust
2.179920e+03 7.772966e-04 1.354703e-03 7.702404e-04 7.578092e-04

> # The estimators have (estimated) MSE:
> est.mse

        IPTW   modifiedHT Subtract1000 SubtractMean DoubleRobust
2.180170e+03 7.773441e-04 1.354808e-03 7.702405e-04 7.578103e-04
```

Solution:

Appendix A: a specific data generating process

The following code was used to generate the data sets `RAssign3.csv` and `RAssign3.missing.cvs`. In this data generating process (one of many compatible with the structural causal model $\mathcal{M}^*$), all background errors are independent.

```
> #-----------
> # genData: function to generate the data
> # input: sample size n
```

```
> # output: data frame with W1, W2, W3, W4, A, Delta, Y, as well as counterfactual outcomes
> #-------
> genData<- function(n){
+   U.Y <- runif(n, 0, 1)
+   W1 <- rbinom(n, size=1, prob=.5) # urban/rural
+   W2 <- round(runif(n, 50, 75)) # age
+   W3 <- round(rnorm(n), 2) # centered measure of CVD
+   W4 <- round(rnorm(n), 2) # centered measure of SES
+   pscore <- plogis(1 - W1 - W2/40 + 1.5*W3 + 1.5*W4)
+   #hist(pscore)
+   A <- rbinom(n, size =1, prob= pscore) # own dog or not
+   # generate measurement indicator
+   pdelta <- plogis(3+ 2*A +W1 - W2/40 - 2*W3 + .5*W4)
+   #hist(pdelta)
+   Delta <- rbinom(n, size=1, prob=pdelta)
+
+   get.outcome <- function(W1, W2, W3, W4, A, U.Y){
+    # generate outcome under complete measurement
+     pdie <- plogis(-3 - 2.5*A -2*W1 + 2*W2/40 -W3 - W4 )
+     as.numeric( U.Y < pdie )
+   }
+   Y.1<- get.outcome(W1, W2, W3, W4, A=1, U.Y)
+   Y.0<- get.outcome(W1, W2, W3, W4, A=0, U.Y)
+   # outcome if no missingness
+   Y <- get.outcome(W1, W2, W3, W4, A, U.Y)
+   # outcome if allow for missingess
+   Y.na <- Y
+   Y.na[Delta==0] <- 0
+   data.frame(W1, W2, W3, W4, A, Y.1, Y.0, Delta,  Y, Y.na)
+ }

> # create the RAssign3.csv & RAssign3.missing.csv
> set.seed(690)
> Full<- genData(n=2500)
> ObsData<- subset(Full, select=c(W1, W2, W3, W4, A, Y))
> write.csv(ObsData, file="RAssign3.csv", row.names=F)
> ObsData<- subset(Full, select=c(W1, W2, W3, W4, A, Delta, Y.na))
> colnames(ObsData) <- c('W1', 'W2', 'W3', 'W4', 'A', 'Delta','Y')
> write.csv(ObsData, file="RAssign3.missing.csv", row.names=F)
```

• Given this specific data generating process, we could estimate the causal effect drawing a huge number of observations and taking the average contrast in the counterfactual outcomes.

```
> set.seed(690)
> # calculate true ATE by drawing a huge number of observations
> nTot=100000
> Full <- genData(n=nTot)
> # causal risk difference
> mean(Full$Y.1) - mean(Full$Y.0)

[1] -0.27283

> # causal risk ratio
> mean(Full$Y.1) / mean(Full$Y.0)
```

```
[1] 0.2468877
```

The counterfactual risk of mortality would $\Psi^*(\mathbb{P}^*)$=26.4% higher if all older adults with cardiovascular disease owned a dog versus did not! #DogsAndDAGs