R Lab 4 - Estimation, Part I: IPTW

Advanced Topics in Causal Inference

Assigned: October 6, 2020
Lab due: October 13, 2020 on bCourses. 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. Upload your own completed lab to bCourses.

Last lab:
Make explicit assumptions sufficient for identifiability, and under those assumptions, write the target causal parameter as a statistical parameter.

This lab:
1. IPTW estimation for simple treatment specific means.
2. IPTW estimation for working marginal structure models.
3. Evaluate performance of IPTW estimators.

Next lab:
Introduction to the `ltmle` package.

---

1 Introduction and Motivation

Last lab we derived statistical parameters (functions of the observed data distribution, P_0) that, under certain assumptions, are equal to our target causal parameters (functions of the counterfactual distribution, $P_{U,X}$) of interest. These statistical parameters are also called *estimands*; specifically, in the last lab we derived g-computation estimands. In this lab, we will introduce another statistical parameter: the IPTW estimand. It turns out that g-computation and IPTW estimands are equivalent!

1.1 Showing the g-computation and IPTW estimands are equivalent

Suppose we have $t = 1, 2$ timepoints, no censoring, and we want to estimate the causal parameter $E[Y_{\bar{a}(2)^*}]$.

Under sequential randomization and positivity, the corresponding IPTW estimand is equal to:

$$\Psi(P_0) = E_0\Big[\frac{\mathbb{I}[\bar{A}(2) = \bar{a}(2)^*]}{g_0(A(1)|L(1))g_0(A(2)|\bar{L}(2), A(1))}Y\Big] \tag{1}$$

Using the definition of expectations, (e.g., $E[X] = \sum_x xP(X = x)$) equation (1) can be rewritten as:

$$
\begin{aligned}
&= \sum_{y,\bar{a}(2),\bar{l}(2)} y \\
&\quad \times P_0(Y = y, \bar{A}(2) = \bar{a}(2), \bar{L}(2) = \bar{l}(2)) \\
&\quad \times \frac{\mathbb{I}[\bar{A}(2) = \bar{a}(2)^*]}{g_0(A(1) = a(1)|L(1) = l(1))g_0(A(2) = a(2)|\bar{L}(2) = \bar{l}(2), A(1) = a(1))}
\end{aligned} \tag{2}
$$

Factorize $P_0(Y = y, \bar{A}(2) = \bar{a}(2), \bar{L}(2) = \bar{l}(2))$ and re-write equation (2) as:

$$
\begin{aligned}
&= \sum_{y,\bar{l}(2)} y \\
&\quad \times P_0(Y = y|\bar{A}(2) = \bar{a}(2)^*, \bar{L}(2) = \bar{l}(2)) \\
&\quad \times P_0(A(2) = a(2)^*|\bar{L}(2) = \bar{l}(2), A(1) = a(1)^*) \\
&\quad \times P_0(L(2) = l(2)|A(1) = a(1)^*, L(1) = l(1)) \\
&\quad \times P_0(A(1) = a(1)^*|L(1) = l(1)) \\
&\quad \times P_0(L(1) = l(1)) \\
&\quad \times \frac{1}{g_0(A(1) = a(1)^*|L(1) = l(1))g_0(A(2) = a(2)^*|\bar{L}(2) = \bar{l}(2), A(1) = a(1)^*)}
\end{aligned} \tag{3}
$$

Canceling out terms (remember that $P_0(A|L) = g_0(A|L)$), we can rewrite equation (3) as:

$$
\begin{aligned}
&= \sum_{y,\bar{l}(2)} y \\
&\quad \times P_0(Y = y|\bar{A}(2) = \bar{a}(2)^*, \bar{L}(2) = \bar{l}(2)) \\
&\quad \times P_0(L(2) = l(2)|A(1) = a(1)^*, L(1) = l(1)) \\
&\quad \times P_0(L(1) = l(1))
\end{aligned} \tag{4}
$$

Again, by the definition of expectations:

$$
\begin{aligned}
&= \sum_{\bar{l}(2)} E[Y|\bar{A}(2) = a(2)^*, \bar{L}(2) = \bar{l}(2)] \\
&\quad \times P_0(L(2) = l(2)|A(1) = a(1)^*, L(1) = l(1)) \\
&\quad \times P_0(L(1) = l(1))
\end{aligned} \tag{5}
$$

The last equation (5) is exactly the longitudinal g-computation formula! (Recall that we derived this g-computation estimand in `R` Lab 3).

1.2 Estimation using IPTW and modified IPTW

Now onto estimation!

In previous labs, we had been dealing with distributions (e.g., P_0) and values of functions of these distributions (e.g., $\Psi(P_0)$, or statistical estimands) that are often unknown to us when dealing with applied data. Now, we are handed a dataset – a *finite* sample of n i.i.d. observations of O, which follows the empirical distribution, P_n. P_n puts weight $\frac{1}{n}$ on each observation, O_i, $i = 1, ..., n$. Using a function of the finite sample drawn from P_n, we aim to estimate our statistical estimands.

Recall that the IPTW *estimand* is the following (note this is equation (1) above, extended to K timepoints):

$$\Psi(P_0) = E\left[\frac{\mathbb{I}[\bar{A}(K) = \bar{a}(K)]}{\prod_{t=1}^{K} g_0(A(t)|\bar{A}(t-1), \bar{L}(t))} Y\right]$$

where $g_0(A(t)|\bar{A}(t-1), \bar{L}(t))$ is the product of time-point-specific predicted probabilities of observed treatment and/or censoring, given observed treatment and covariate history.

The **IPTW *estimator*** (that is aimed at estimating the IPTW estimand) applied to our empirical data is then:

$$\begin{aligned}\hat{\Psi}(P_n) &= \frac{1}{n}\sum_{i=1}^{n} \frac{\mathbb{I}[\bar{A}_i(K) = \bar{a}_i(K)]}{\prod_{t=1}^{K} g_n(A_i(t)|\bar{A}_i(t-1), \bar{L}_i(t))} Y_i \\ &= \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}[\bar{A}_i(K) = a_i(K)]\hat{w}_i Y_i\end{aligned}$$

where $\hat{w}_i = \frac{1}{g_n(A_i(t)|\bar{A}_i(t-1), \bar{L}_i(t))}$ (i.e., the IPTW weights).

We will also implement a modified version of the IPTW estimator, called the **modified Horvitz-Thompson estimator**:

$$\begin{aligned}\hat{\Psi}(P_n) &= \frac{\sum_{i=1}^{n} \frac{\mathbb{I}[\bar{A}_i = \bar{a}]}{\prod_{t=1}^{K} g_n(A_i(t)|\bar{L}_i(t), \bar{A}_i(t-1))} Y_i}{\sum_{i=1}^{n} \frac{\mathbb{I}[\bar{A}_i = \bar{a}]}{\prod_{t=1}^{K} g_n(A_i(t)|\bar{L}_i(t), \bar{A}_i(t-1))}} \\ &= \frac{\sum_{i=1}^{n} \mathbb{I}[\bar{A}_i(K) = \bar{a}_i(K)]\hat{w}_i Y_i}{\sum_{i=1}^{n} \mathbb{I}[\bar{A}_i(K) = \bar{a}(K)]\hat{w}_i}\end{aligned}$$

This is essentially the standard IPTW estimator, but we divide by the sample average of the weights. The advantage to the Horvitz-Thompson estimator is that it 1) may reduce the variability of the IPTW estimates and 2) respects the parameter space (for example, if Y is binary, we will ensure that we don't get an expectation less than 0 or greater than 1).

1.3 Using IPTW weights to estimate MSM parameters

Let $m(\bar{a}|\beta)$ be a (working) model for $E[Y_{\bar{a}}]$. For example, our working MSM could be:

$$m(\bar{a}|\beta) = \beta_0 + \beta_1 \sum_{t=1}^{K} a(t)$$

Then, if we either assume the MSM is correctly specified or if we choose $g^*(\bar{A}) = 1$ as the projection function, then our target causal parameter is:

$$\Psi^F(P_{U,X}) = \arg\min_{\beta} \sum_{\bar{a}\in\mathcal{A}} (Y_{\bar{a}} - m(\bar{a}|\beta))^2$$

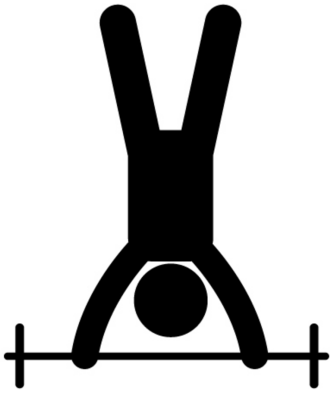

Figure 1: Inverse weighting.

We can estimate this parameter (the βs) using a weighted regression with unstabilized weights (which are just the weights, $\hat{w}_i$, above):

$$\hat{w}_i = \frac{1}{\prod_{t=1}^{K} g_n(A_i(t)|\bar{A}_i(t-1), \bar{L}_i(t))}$$

We could alternatively define our target parameter value β using a different projection function. For example, a common choice is $g(\bar{A})$, defined as the marginal probability of regime $\bar{A}$ occurring. Now the target causal parameter is defined as:

$$\Psi^F(P_{U,X}) = \arg\min_{\beta} \sum_{\bar{a} \in \mathcal{A}} (Y_{\bar{a}} - m(\bar{a}|\beta))^2 g(\bar{A})$$

If the MSM is correctly specified, the choice of projection function $g(\bar{A})$ will not change the value of β that minimizes the sum of the squared residuals (and therefore will not change the true value of the target causal parameter). However, if the MSM is not correctly specified, the choice of projection function will result in a different true value of the target causal parameter – one that puts more weight on the treatment regimes with more representation in the data.

For this second target parameter, an IPTW estimator of β can be implemented using the following stabilized weights:

$$s\hat{w}_i = \frac{g_n(\bar{A}_i(K))}{\prod_{t=1}^{K} g_n(A_i(t)|\bar{A}_i(t-1), \bar{L}_i(t))}$$

The numerator, $g_n(\bar{A}_i(K))$, is the marginal probability of seeing the subject's particular treatment regime. Informally, the advantage to using stabilized weights is that it prevents rare exposure histories from being assigned large weights.

2 This lab

Finally, your GSR gives you some data! One-thousand students were sampled from the school, which we can treat as 1,000 i.i.d. copies of O.

This lab will walk you through loading your GSR data, estimating (using IPTW) the parameters that inform us about the effects of sleep, and evaluating how well IPTW works. IPTW performance will be measured by evaluating the estimator's bias (or, the expected difference between the point estimate, ψ_n, and the true parameter value):

$$Bias(\hat{\Psi}(P_n)) = E_0[\hat{\Psi}(P_n) - \Psi(P_0)]$$

the estimator's variance (or, the average squared difference between each estimate and the average of the estimates):

$$Variance(\hat{\Psi}(P_n)) = E_0\left[\left[\hat{\Psi}(P_n) - E_0[\hat{\Psi}(P_n)]\right]^2\right]$$

and the estimator's mean-squared error (MSE, or, the average squared distance the estimator is from the true parameter value):

$$E_0\left[[\hat{\Psi}(P_n) - \Psi(P_0)]^2\right] = Bias^2 + Variance$$

For this lab, you'll answer questions for two of the data generating systems we've been working with in previous labs. We are now on step 6 (estimation) of the causal roadmap. See `R` Lab 1 for specifying the causal model (step 1), `R` Lab 2 for the causal questions and parameters of interest (step 2), `R` Lab 1 and 2 for the link between the SCM and observed data (step 3), and `R` Lab 3 for identifiability and specification of the target parameter of the observed data distribution (step 4 and 5, respectively).

2.1 To turn in:

For each of the 2 data structures listed below, answer the following questions:

1. **Implement IPTW for estimation** of statistical parameters that, under sequential randomization and positivity assumptions, are equal to our causal parameters of interest:
   - For both data structures, implement IPTW for estimation of treatment specific means. Implement both the **standard IPTW** and **modified Horvitz-Thompson IPTW** estimators.
   - For Data Structure 2 ONLY, implement IPTW to estimate the parameters of an MSM using an **MSM with unstabilized weights**. As a bonus, you may additionally estimate the parameters of an MSM with stabilized weights. (Fall 2020 – not required)

   Within the IPTW estimation process, **comment on the distributions of the predicted probabilities and weights** that went into the IPTW estimates. Also, **interpret each IPTW estimate.**
2. **Compare performance metrics of each of the estimators**. Specifically, evaluate the bias, variance, and mean-squared error (MSE) of each of the estimators.

Data Structure 0: $O = (L(1), A(1), L(2), A(2), Y)$

Target parameter: we are interested in the expected Y if everyone got treatment regime $\bar{a}(2) = 1$.

$$\Psi^F(P_{U,X}) = E_{U,X}[Y_{\bar{a}(2)=1}]$$

1. **IPTW for estimation:**

   (a) Load `DataStructure0.RData` using the `load()` function. Make sure you have specified the correct file path. You should see 3 new things come up in your global environment:

   - `ObsData0` – this the data given to you by your GSR. It is a dataframe of 1,000 observations (note: it follows Data Structure 0 from the previous lab).
   - `Psi.F0` – this is the true $\Psi^F(P_{U,X})$ value for the target causal parameter $E_{U,X}[Y_{\bar{a}(2)=1}]$ (generated in lab 3).
   - `generate_data0` – this is the function that generates n copies of Data Structure 0.
   - `generate_data0_intervene` – we won't use this function in this lab, so if you'd like you can remove it from your global environment using the `rm()` function.

     ```
     > rm(generate_data0_intervene)
     ```

   (b) Assign the number of students to `n`.

   (c) Estimate the IPTW weights:

   i. Estimate the probability of receiving treatment $P_0(A(t) = 1|\bar{L}(t), \bar{A}(t-1)) = g_0(A(t) = 1|\bar{L}(t), \bar{A}(t-1))$ for $t = 1, 2$ using correctly specified parametric regression models. The correct model specifications are (optional: refer back to `R` lab 3 to verify!):

   $$g_0(A(1) = 1|L(1)) = expit[\beta_0 + \beta_1 L(1)]$$
   $$g_0(A(2) = 1|\bar{L}(2), A(1)) = expit[\beta_0 + \beta_1 L(1) + \beta_2 L(2)]$$

   Use the `glm()` function, and specify the arguments `family = 'binomial'` for logistic regression and `data = ObsData0` to call the correct dataset.

   ii. Predict each subject's probability of the exposure at time t, given his or her observed exposure and covariate history, i.e., $g_n(A_i(t) = 1|\bar{A}_i(t-1), \bar{L}_i(t))$. Name these vectors `gA1.1` and `gA2.1`. Also evaluate each subject's probability of the observed exposure at time t, given their observed exposure and covariate history, i.e., $g_n(A_i(t)|\bar{A}_i(t-1), \bar{L}_i(t))$. Name these vectors `gA1` and `gA2`. *Hint:* use the `predict()` function on each of the logistic regressions applied above, specifying `type = 'response'` as an argument.

   iii. **Look at the distributions of the predicted probabilities of exposure using the `hist()` and `summary()` functions. What are we able to assess by looking at the distribution of the predicted probabilities of exposure? Any cause for concern here?**

   iv. Obtain the indicator variable $\mathbb{I}[\bar{A}_i(2) = 1]$ by creating a logical variable that indicates which students had a treatment history $\bar{a}(2) = 1$:

   ```
   > I11 = ObsData0$A1 == 1 & ObsData0$A2 == 1
   ```

   v. Calculate the weights $\hat{w}_i$ by taking the inverse of the product of the time point specific predicted probabilities of the observed treatment, i.e.:

   $$\hat{w}_i = \frac{1}{g_n(A_i(1)|L_i(1)) \times g_n(A_i(2)|A_i(1), L_i(1), L_i(2))}$$

   Do this *only for people who got treatment history* $\bar{a} = 1$ by using the indicator variable you created in the previous step. **Look at the distribution of these weights. What can we assess by looking at the distribution of the estimated weights? Any cause for concern here? *Bonus*: what are the drawbacks of only using this method of assessment, as opposed to looking at the propensity scores (as in step two steps ago)?**

(d) Implement the **standard IPTW estimator** by taking the empirical mean of the weighted outcomes:

$$\hat{\Psi}(P_n) = \frac{1}{n}\sum_{i=1}^{n} \frac{\mathbb{I}[\bar{A}_i(2)=1]}{g_n(A_i(1)|L_i(1)) \times g_n(A_i(2)|A_i(1), L_i(1), L_i(2))} Y_i$$

$$= \frac{1}{n}\sum_{i=1}^{n} \underbrace{\mathbb{I}[\bar{A}_i(2)=1]}_{\textit{You obtained this in step iv!}} \times \underbrace{\hat{w}_i}_{\textit{You obtained this in step v!}} \times \underbrace{Y_i}_{\textit{This is the outcome variable in the data!}}$$

(e) Implement the **modified Horvitz-Thompson estimator**:

$$= \frac{\frac{1}{n}\sum_{i=1}^{n} \mathbb{I}[\bar{A}(2)_i = 1]\hat{w}_i Y_i}{\frac{1}{n}\sum_{i=1}^{n} \mathbb{I}[\bar{A}(2)_i = 1]\hat{w}_i}$$

where, again, $\hat{w}_i = \frac{1}{g_n(A_i(1)|L_i(1)) \times g_n(A_i(2)|A_i(1), L_i(1), L_i(2))}$.
Hint: Divide the result from the previous step by the average of the IPTW weights!

(f) **Interpret your results.**

2. **Estimator performance metrics:**

(a) Set the number of iterations `B` to 5 (to start).

(b) Create a matrix `estimates_data0` with `B` rows and 2 columns. Name the columns of the matrix `IPTW` and `IPTW.HT`.

(c) Within a for loop from `b` to `1:B`, do the following:

i. Redraw n copies of the data using the `generate_data0()` function you loaded earlier.

ii. Copy and paste code from the previous section (making sure to remove any plots!) to generate the new IPTW estimates using the redrawn data in the previous step. Specifically:

1. Estimate the treatment mechanism
2. Predict the conditional probability of having the exposure
3. Create weights
4. Generate the standard and modified Horvitz-Thompson IPTW estimates

iii. Save the estimates in the b^{th} row of the `estimates_data0` matrix.

```
> estimates_data0[b,] = c(IPTW, IPTW.HT)
```

(d) When you are confident that your code is working, set the seed to 252 and increase the number of iterations `B = 500` and rerun your code.

(e) For each estimator, estimate the:

- Bias. *Hint:* use the `colMeans()` function.
- Variance. *Hint:* use the `var()` function on the estimates to get the covariance matrix, and take the diagonal of that matrix using the `diag()` function to get each estimator's variance.
- MSE. *Hint:* use the `colMeans()` function.

Data Structure 2: $O = (L(1), A(1), L(2), A(2), L(3), A(3), L(4), A(4), Y)$

Target parameter 1: we are interested in the difference in the expected test score if all students got 8 or more hours of sleep for all 4 nights before the test versus if all students got less than 8 hours of sleep for all 4 nights before the test.

$$\Psi^F(P_{U,X}) = E_{U,X}[Y_{\bar{a}(4)=1} - Y_{\bar{a}(4)=0}]$$

Target parameter 2: we are also interested in how the expectation of counterfactual statistics test scores varies as a function of total nights on which a student got more than 8 hours of sleep.

$$m(\bar{a}|\beta) = E[Y_{\bar{a}}] = \beta_0 + \beta_1 \sum_{t=1}^{4} a(t)$$

Specifically, ψ^F is the true value of β_1 (and if the MSM is not correctly specified, then it is the true value defined according to some projection, either 1 or $g(\bar{A})$). In particular, β_1 asks: for one additional night of 8 or more hours of sleep, what is the change in students' mean counterfactual test score?

1. **IPTW for estimation:**

   (a) Load `DataStructure2.RData` using the `load()` function. Make sure you have specified the correct file path. You should see 5 new things come up in your global environment:
   - `ObsData2` – this is a dataframe of 1,000 students that follows Data Structure 2 from previous labs.
   - `Psi.F2` – this is the true $\Psi^F(P_{U,X})$ value for the target causal parameter $E_{U,X}[Y_{\bar{a}(4)=1}] - E_{U,X}[Y_{\bar{a}(4)=0}]$ (generated in lab 2).
   - `TrueMSMbeta1` – this is the true $\Psi^F(P_{U,X})$ value for the target causal parameter β_1 from $m(\bar{a}|\beta)$, our MSM.
   - `TrueMSMbeta1_wts` – this is the true $\Psi^F(P_{U,X})$ value for the target causal parameter β_1 from $m(\bar{a}|\beta)$, our weighted MSM.
   - `generate_data2` – this is the function that generates n copies of Data Structure 2.
   - `generate_data2_intervene` – we won't use this function in this lab.

   (b) Assign the number of students to `n`.

   (c) Estimate the IPTW weights:

   i. Estimate the probability of receiving treatment $P_0(A(t) = 1|\bar{L}(t), \bar{A}(t-1)) = g_0(A(t) = 1|\bar{L}(t), \bar{A}(t-1))$ for $t = 1, ..., 4$ using correctly specified parametric regression models. This is the conditional probability of getting 8 or more hours of sleep at time t, given the student's treatment and covariate history. The correct model specifications are (optional: refer back to `R` lab 1 to verify!):

$$g_0(A(1) = 1|L(1)) = expit[\beta_0 + \beta_1 L(1)]$$
$$g_0(A(2) = 1|\bar{L}(2), A(1)) = expit[\beta_0 + \beta_1 L(1) + \beta_2 A(1) + \beta_3 L(2)]$$
$$g_0(A(3) = 1|\bar{L}(3), \bar{A}(2)) = expit[\beta_0 + \beta_1 L(1) + \beta_2 A(1) + \beta_3 L(2) + \beta_4 A(2) + \beta_5 L(3)]$$
$$g_0(A(4) = 1|\bar{L}(4), \bar{A}(3)) = expit[\beta_0 + \beta_1 L(1) + \beta_2 A(1) + \beta_3 L(2) + \beta_4 A(2) + \beta_5 L(3) + \beta_6 A(3) + \beta_7 L(4)]$$

   Use the `glm()` function, and specify the arguments `family = 'binomial'` for logistic regression and `data = ObsData2`.

   ii. Predict each student's probability of the exposure at time t, given previous exposure and covariate history: $g_n(A_i(t) = 1|\bar{A}_i(t-1), \bar{L}_i(t))$.

   A. Obtain the predicted probabilities of getting 8 or more hours of sleep for each timepoint, and assign to the variables `gA1.1, gA2.1, gA3.1, gA4.1`.
   Hint: use the `predict()` function on each of the logistic regressions applied above, specifying `type = 'response'` as an argument.

B. Look at the distributions of the predicted probabilities of getting 8 or more hours of sleep at time t, given the past, using the `hist()` and `summary()` functions. **Comment on the distributions. Any cause for concern?**

iii. Obtain the observed conditional probabilities of getting 8 or more hours of sleep: $g_n(A_i(t)|\bar{A}_i(t-1), \bar{L}_i(t))$. That is, for each timepoint, among students who got 8 or more hours of sleep at that timepoint, assign the predicted conditional probability of getting 8 or more hours of sleep. Similarly, for each timepoint, among students who got less than 8 hours of sleep at that timepoint, assign the predicted conditional probability of getting less than 8 hours of sleep.
Hint: For example, for timepoint 1:

```
> gA1 = (ObsData2$A1 == 1) * gA1.1 + (ObsData2$A1 == 0) * (1 - gA1.1)
> # equivalently
> gA1 = ifelse(ObsData2$A1 == 1, gA1.1, 1-gA1.1)
```

Repeat for $t = 2, 3, 4$.

iv. Calculate the weights, $\hat{w}_i$, for each subject by taking the inverse of the product of the time point specific predicted probabilities:

$$\hat{w}_i = \frac{1}{\prod_{t=1}^{4} g_n(A_i(t)|\bar{A}_i(t-1), \bar{L}_i(t))}$$

Look at the distribution of the weights by using the `hist()` and `summary()` functions. **Comment on the distribution of the weights.**

v. Create a logical variable that indicates which students had a treatment history $\bar{a}(4) = 1$. For example:

```
> I1111 = ObsData2$A1 == 1 & ObsData2$A2 == 1 & ObsData2$A3 == 1 & ObsData2$A4 == 1
```

Repeat for students who had treatment history $\bar{a}(4) = 0$.

vi. Look at the distribution of the weights for students who had $\bar{a}(4) = 1$, and then $\bar{a}(4) = 0$. **Comment on the distributions. In particular, what does the distribution of the weights tell you here that the time-point specific conditional probabilities of treatment do not?**

(d) **Evaluating target parameter 1:**

i. Implement the **standard IPTW estimator** by taking the empirical mean of the weighted outcomes:

$$\begin{aligned}\hat{\Psi}(P_n) &= \frac{1}{n}\sum_{i=1}^{n} \frac{\mathbb{I}[\bar{A}_i(4) = 1]}{\prod_{t=1}^{4} g_n(A_i(t)|\bar{A}_i(t-1), \bar{L}_i(t))} Y_i - \frac{1}{n}\sum_{i=1}^{n} \frac{\mathbb{I}[\bar{A}_i(4) = 0]}{\prod_{t=1}^{4} g_n(A_i(t)|\bar{A}_i(t-1), \bar{L}_i(t))} Y_i \\ &= \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}[\bar{A}_i(4) = 1]\hat{w}_i Y_i - \sum_{i=1}^{n} \mathbb{I}[\bar{A}_i(4) = 0]\hat{w}_i Y_i\end{aligned}$$

ii. **Interpret your results.**

iii. Implement the **modified Horvitz-Thompson estimator** by dividing by the mean of the IPTW weights:

$$= \frac{\frac{1}{n}\sum_{i=1}^{n} \mathbb{I}[\bar{A}_i(4) = 1]\hat{w}_i Y_i}{\frac{1}{n}\sum_{i=1}^{n} \mathbb{I}[\bar{A}_i(4) = 1]\hat{w}_i} - \frac{\frac{1}{n}\sum_{i=1}^{n} \mathbb{I}[\bar{A}_i(4) = 0]\hat{w}_i Y_i}{\frac{1}{n}\sum_{i=1}^{n} \mathbb{I}[\bar{A}_i(4) = 0]\hat{w}_i}$$

where, again, $\hat{w}_i = \frac{1}{\prod_{t=1}^{4} g_n(A_i(t)|\bar{A}_i(t-1), \bar{L}_i(t))}$.

iv. **Interpret your results.**

(e) **Evaluating target parameter 2 (Fall 2020 – not required)**:

i. Calculate each subject's observed cumulative $\bar{a}$ by using the `rowSums()` function on `A1, A2, A3` and `A4`:

```
> sum.a = rowSums(ObsData2[c("A1", "A2", "A3", "A4")])
```

ii. For each timepoint, estimate the treatment mechanism and predict each subject's probability of sleeping 8 or more hours.
Hint: you already did this! Skip to the next step

iii. **MSM parameter estimation using unstabilized IPTW weights**

A. Create the weight vector `w` as the inverse of the product of the timepoint specific predicted probabilities:

$$\hat{w}_i = \frac{1}{\prod_{t=1}^{4} g_n(A_i(t)|\bar{A}_i(t-1), \bar{L}_i(t))}$$

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

B. Estimate β_1 according to the MSM by running a weighted regression of outcome `Y` on `sum.a`.
Hint: Use the `glm()` function to run the regression, remembering to specify the arguments `weights` and `data`. Note that we do not want to use the standard errors provided here. In order to get inference on the coefficients of an MSM, we would need to use robust standard errors or bootstrap.

C. **Interpret your results.**

iv. **Bonus: MSM parameter estimation using *stabilized* IPTW weights** Note: see `R` Lab 2 solutions for a calculation of the true causal parameter value defined using a working MSM with a projection function $g^*(\bar{A})$ other than 1.

A. Evalute the weights, $s\hat{w}_i$ as the weighted inverse of the product of the timepoint specific predicted probabilities:

$$s\hat{w}_i = \frac{g^*(\bar{A}_i(t))}{\prod_{t=1}^{4} g_n(A_i(t)|\bar{A}_i(t-1), \bar{L}_i(t))}, \text{ where } g^*(\bar{A}_i(t)) = g_n(\bar{A}_i(4))$$

1. Calculate the denominator $\prod_{t=1}^{4} g_n(A_i(t)|\bar{A}_i(t-1), \bar{L}_i(t))$
Hint: you already did this! Skip to the next step.
2. Calculate the numerator $g(\bar{A}_i(4))$. In other words, these are the marginal probabilities of each of the 16 treatment regimes.

(a) Create every permutation of $\bar{A}(4)$:

```
> abar = expand.grid(c(0,1), c(0,1), c(0,1), c(0,1))
```

(b) Create a new vector `g.abar` of `NA`s of length n to store the marginal probability of the $\bar{a}(4)$ associated with each subject.

(c) For each treatment regime, find the marginal probability of observing that particular $\bar{a}(4)$ (e.g., the proportion of that $\bar{a}(4)$) and assign it to the subjects for whom we observe that $\bar{a}(4)$.
Hint: Use the following `for` loop:

```
> for(i in 1:16){
+   # marginal probability
+   marg.prob = mean(ObsData2$A1 == abar[i,1] &
+                      ObsData2$A2 == abar[i,2] &
+                      ObsData2$A3 == abar[i,3] &
+                      ObsData2$A4 == abar[i,4])
+   # assign to subject in vector g.abar
+   g.abar[ObsData2$A1 == abar[i,1] &
+            ObsData2$A2 == abar[i,2] &
+            ObsData2$A3 == abar[i,3] &
+            ObsData2$A4 == abar[i,4]] = marg.prob
+ }
```

B. Multiply the original weights, `w`, by the numerator, `g.abar`, to obtain each subject's stabilized weight, $s\hat{w}_i$.

C. Examine the distribution of the stabilized weights, and compare with the unstabilized weights.

D. Estimate β_1 according to the MSM by running a weighted regression of outcome `Y` on `sum.a`. *Hint:* Use the `glm()` function to run the regression, remembering to specify the arguments `weights` and `data`.

E. Interpret your results.

Solution:

```
> # calculate the observed cumulative abar(4) for each subject
> sum.a = rowSums(ObsData2[c("A1", "A2", "A3", "A4")])

> # estimate the parameters of the MSM with IPTW weights
> IPTW.MSM = glm(Y ~ sum.a, data = ObsData2, weights = w)
> IPTW.MSM.coef = IPTW.MSM$coefficients[["sum.a"]] # extract the Beta1 coefficient
> IPTW.MSM.coef

[1] 3.521134
```

The estimated parameters of the MSM are

$$\begin{aligned} m(\bar{a}|\beta) &= \hat{\beta}_0 + \hat{\beta}_1 \sum_{t=1}^{4} a(t) \\ &= 57.38 + 3.52 \sum_{t=1}^{4} a(t) \end{aligned}$$

Under the necessary causal assumptions, and if we assume the MSM is correct, we can interpret $\hat{\beta}_1$ as follows: for one more night of sleep, the average test score increases by 3.52. Recall that the true value of β_1 using projection function $g^*(\bar{A}) = 1$ was 3.4.

Bonus! Calculating the parameters of an MSM using *stabilized* IPTW weights.

```
> # create every permutation of abar(4)
> abar = expand.grid(c(0,1), c(0,1), c(0,1), c(0,1))

> # create a new vector for the marginal probability of each permutation of abar(4)
> # in other words, g(abar(4))
> g.abar = rep(NA, n)

> # for each possible permutation of abar(4), find the marginal probability.
> # assign that probability to each subject who has that observed abar(4).
> for(i in 1:16){
+
+   marg.prob = mean(ObsData2$A1 == abar[i,1] &
+                      ObsData2$A2 == abar[i,2] &
+                      ObsData2$A3 == abar[i,3] &
+                      ObsData2$A4 == abar[i,4]) # marginal probability
+
+   g.abar[ObsData2$A1 == abar[i,1] &
+            ObsData2$A2 == abar[i,2] &
+            ObsData2$A3 == abar[i,3] &
+            ObsData2$A4 == abar[i,4]] = marg.prob # assign to subject in vector g.abar
+
+ }
```

```
> # multiply the original weights by the numerator to obtain the the stabilized weights
> sw = g.abar*w

> summary(sw)

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.6745  0.9091  1.0076  1.0112  1.0961  1.6019

> hist(sw, main = "Distribution of stabilized weights")
```

After stabilization, the weights are less variable and extreme.

```
> # estimate the parameters of the MSM with IPTW stabilized weights
> IPTW.MSM.stab = glm(Y ~ sum.a, data = ObsData2, weights = sw)
> IPTW.MSM.stab.coef = IPTW.MSM.stab$coefficients[[2]] # extract the Beta1 coefficient
> IPTW.MSM.stab.coef

[1] 3.482513
```

The estimated parameters of the MSM using IPTW with stabilized weights are

$$\begin{aligned} m(\bar{a}|\beta) &= \hat{\beta}_0 + \hat{\beta}_1 \sum_{t=1}^{4} a(t) \\ &= 57.47 + 3.48 \sum_{t=1}^{4} a(t) \end{aligned}$$

Recall that the β_1 on our true causal curve using projection function $g^*(\bar{A}) = g(\bar{A}(4))$ is equal to 3.4.

2. **Estimator performance metrics:**

   (a) Set the number of iterations `B` to 5 (to start)

   (b) Create a matrix `estimates_data2` with `B` rows and 2 columns. Name the columns of the matrix `IPTW, IPTW.HT`.
   Note: If you calculated the MSM parameters without and with stabilized weights (i.e., the bonus question in the previous section), create 4 columns in your matrix instead of 2, and name the 3^{rd} and 4^{th} columns `IPTW.MSM.coef` and `IPTW.MSM.stab.coef`, respectively.

   (c) Within a for loop from `b` to `1:B`, do the following:

      i. Redraw n copies of the data using the `generate_data2()` function you loaded earlier.

      ii. Copy and paste code from the previous steps (making sure to remove any plots!) to generate 2 new IPTW estimates (4 IPTW estimates if you calculated the MSM parameters) using the redrawn data in the previous step. Specifically:
         1. Estimate the treatment mechanism
         2. Predict the conditional probability of having the exposure
         3. Predict the probability of having the observed exposure
         4. Create weights
         5. Generate the standard and Horvitz-Thompson IPTW estimates
         6. Generate coefficient estimates of the MSM (Fall 2020 – not required)

      iii. Save the estimates in the b^{th} row of the `estimates_data2` matrix.

```
> estimates_data2[b,] = c(IPTW, IPTW.HT)
> estimates_data2[b,] = c(IPTW, IPTW.HT, IPTW.MSM.coef, IPTW.MSM.stab.coef) # with MSM
```

Note: again, if you estimated a stabilized MSM, add `IPTW.MSM.stab.coef` to the end of this vector.

(d) When you are confident that your code is working, set the seed to 252 and increase the number of iterations `B = 500` and rerun your code.

(e) For each estimator, estimate the:

- Bias. *Hint:* use the `colMeans()` function.
- Variance. *Hint:* use the `var()` function on the estimates to get the covariance matrix, and take the diagonal of that matrix using the `diag()` function to get each estimator's variance.
- MSE. *Hint:* use the `colMeans()` function.

Solution:

```
> ### IPTW estimator performance metrics - Data 2 ###
> # set seed
> set.seed(252)
> # number of iterations
> B = 500


> # matrix to store IPTW estimates
> estimates_data2 = matrix(NA, nrow = B, ncol = 2)
> # column names for matrix
> colnames(estimates_data2) = c("IPTW.MSM.coef", "IPTW.MSM.stab.coef")


> # for loop that creates many iterations of ObsData2 and implements above IPTW estimates
> for(b in 1:B) {
+
+   # redraw the data
+   ObsData2 = generate_data2(n)
+
+   # estimate treatment mechanisms
+   gA1.reg = glm(A1 ~ L1, family = "binomial", data = ObsData2)
+   gA2.reg = glm(A2 ~ L1 + A1 + L2, family = "binomial", data = ObsData2)
+   gA3.reg = glm(A3 ~ L1 + A1 + L2 + A2 + L3, family = "binomial", data = ObsData2)
+   gA4.reg = glm(A4 ~ L1 + A1 + L2 + A2 + L3 + A3 + L4, family = "binomial", data = ObsData2)
+
+   # predicted probability of having exposure, given history
+   gA1.1 = predict(gA1.reg, type = "response")
+   gA2.1 = predict(gA2.reg, type = "response")
+   gA3.1 = predict(gA3.reg, type = "response")
+   gA4.1 = predict(gA4.reg, type = "response")
+
+   # predicted probability of observed exposure, given history
+   gA1 = (ObsData2$A1 == 1) * gA1.1 + (ObsData2$A1 == 0) * (1 - gA1.1)
+   gA2 = (ObsData2$A2 == 1) * gA2.1 + (ObsData2$A2 == 0) * (1 - gA2.1)
+   gA3 = (ObsData2$A3 == 1) * gA3.1 + (ObsData2$A3 == 0) * (1 - gA3.1)
+   gA4 = (ObsData2$A4 == 1) * gA4.1 + (ObsData2$A4 == 0) * (1 - gA4.1)
+
+   # weight = inverse of predicted probability
+   w = 1/(gA1 * gA2 * gA3 * gA4)
+
```

```
+   # get cumulative abars
+   ObsData2$sum.a = rowSums(ObsData2[c("A1", "A2", "A3", "A4")])
+
+   # MSM and coefficient (unstabilized)
+   IPTW.MSM = glm(Y ~ sum.a, data = ObsData2, weights = w)
+   IPTW.MSM.coef = IPTW.MSM$coefficients[[2]]
+
+   # create stabilized weights
+   abar = expand.grid(c(0,1), c(0,1), c(0,1), c(0,1))
+   g.abar = rep(NA, n)
+   for(i in 1:16){
+     marg.prob = mean(ObsData2$A1 == abar[i,1] &
+                        ObsData2$A2 == abar[i,2] &
+                        ObsData2$A3 == abar[i,3] &
+                        ObsData2$A4 == abar[i,4])
+     g.abar[ObsData2$A1 == abar[i,1] &
+              ObsData2$A2 == abar[i,2] &
+              ObsData2$A3 == abar[i,3] &
+              ObsData2$A4 == abar[i,4]] = marg.prob
+   }
+   sw = g.abar*w # stabilized weights
+
+   # MSM and coefficient (stabilized)
+   IPTW.MSM.stab = glm(Y ~ sum.a, data = ObsData2, weights = sw)
+   IPTW.MSM.stab.coef = IPTW.MSM.stab$coefficients[[2]]
+
+   # store estimates in matrix
+   estimates_data2[b,] = c(IPTW.MSM.coef, IPTW.MSM.stab.coef)
+
+ }


> # Bias
> # bias for betas on stabilized and unstabilized MSM
> mean(estimates_data2[,1] - TrueMSMbeta1)


[1] -0.01069938


> mean(estimates_data2[,2] - TrueMSMbeta1_wts)


[1] -0.006618977


> # Variance
> diag(var(estimates_data2))


     IPTW.MSM.coef IPTW.MSM.stab.coef
        0.01553514         0.01705657


> # MSE
> # MSE for beta1 on stabilized and unstabilized MSM
> mean((estimates_data2[,1] - TrueMSMbeta1)^2)
```

```
[1] 0.01561854

> mean((estimates_data2[,2] - TrueMSMbeta1_wts)^2)

[1] 0.01706627
```

3 For Your Project: IPTW Estimation

Think through the following questions and apply them to the dataset you will use for your final project.

1. **Treatment/censoring mechanism**
   (a) What are the necessary parts of your P_0 that you need for IPTW implementation?
   (b) Estimate these parts using parametric regressions or SuperLearner. What do the predicted probabilities of treatment/censoring look like? Any cause for concern?
   (c) Create weights (and stabilized weights if working with an MSM) based on these estimates. What do they look like? Any cause for concern? Was there something you could see in the distribution of the weights that you couldn't spot by just looking at the predicted probabilities in the previous step?
2. **Implement IPTW**
   (a) Implement standard IPTW and modified Horvitz-Thompson IPTW for treatment-specific mean causal questions. Depending on whether you are using an MSM to define your target parameter, use a weighted regression with stablized and unstabilized IPTW weights for evaluating parameters of an MSM.
   (b) Interpret the estimates generated in the previous question in the context of your study.
3. **Performance of IPTW**
   (a) Evaluate the bias, variance, and MSE of your IPTW estimator by comparing it against the "true" causal estimand you generated in `R` Lab 2.

4 Feedback

Please attach responses to these questions to your lab. Thank you in advance!

1. Did you catch any errors in this lab? If so, where?
2. What did you learn in this lab?
3. Do you think that this lab met the goals listed at the beginning?
4. What else would you have liked to review? What would have helped your understanding?
5. Any other feedback?