R Lab 2 - Calculating True Values of Target Causal Parameters Under Longitudinal Interventions

Advanced Topics in Causal Inference

Assigned: September 14, 2021
Lab due: September 21, 2021 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:
1. Explore different data structures, inspired by real-world experiments.
2. Simulate data generating systems that give rise to data we observe.
3. Think of studies in which these data generating systems may occur.

Goals for this lab:
Translate causal questions into target causal parameters and intervene on the Structural Causal Models (SCMs) to evaluate them.

Next lab:
Understand time dependent confounding and identifiability in a longitudinal context.

---

1 Introduction and Motivation

In this lab, we're going to play with the "root" of where data comes from to generate hypothetical outcomes that answer our causal questions.

In the previous lab, we learned how to simulate data generating systems that give rise to data we observe. Now, we're going to intervene on those data generating systems by deterministically setting certain variables to constant values (what kind of regime does this correspond to?), according to our ideal experiment/causal question of interest. After intervening, we'll generate many counterfactual outcomes. Then, we can apply a function to the distribution of those counterfactual outcomes, or more generally, to the post-intervention distribution of the data, (the function being $\Psi^F(P_{U,X})$) to evaluate the true value of our target causal parameter of interest that answers our causal question.

Note that it is also often possible to evaluate $\Psi^F(P_{U,X})$ analytically and obtain a closed-form solution. But, as mentioned in the previous lab, we can also turn to simulations to obtain answers computationally.

Figure 1: Target parameter.

2 This lab

Recall that your GSR urgently needs you to determine whether lack of sleep is hurting students' academic performance and health. Also remember that we have perfect knowledge of how these outcomes come to be in the world, and more specifically, how they're impacted by sleep, background variables, and random error. In other words, unlike reality, we know the true data generating process.

We also have the power to intervene on these processes. For example, we can "force" all students to get 8 hours of sleep – would students' statistics test scores improve compared to if students got less than 8 hours of sleep? By how much? What about their probability of getting sick? How would the distribution of these outcomes differ if students got 8 or more hours of sleep for multiple nights in a row before the test? In this lab, based on the same 4 data structures as last lab, you'll come up with causal parameters and evaluate their true values (via simulations) to answer these causal questions.

Refer back to `R` Lab 1 for variable definitions and SCMs.

2.1 To turn in:

For each of the 4 data structures listed below, answer the following questions:

Note: For Data Structure 2, there are 2 causal questions. Go through the following steps for **both** causal questions.

1. **Write the causal target parameter that would answer the causal question posed using the data structure presented. What are the counterfactual outcomes?** Explain using notation and in words.
2. **Explain how to intervene on the SCM to get at the causal question/parameter of interest.** Refer back to `R` Lab 1 for each data structure's SCM.
3. **Implement the intervention described in step 2** by updating the data generating function you created in `R` Lab 1.
4. **Evaluate $\Psi^F(P_{U,X})$ via simulated counterfactuals.** Given a large sample of counterfactual outcomes (say, $n = 100,000$), we can closely approximate $\Psi^F(P_{U,X})$.
5. **Write a sentence interpreting the value you got for $\Psi^F(P_{U,X})$.**

Data Structure 1: $O = (W, A, L, \Delta, \Delta Y)$

Causal question: What is the absolute difference in expected test score if all students slept 8 or more hours compared to if all students slept less than 8 hours, under a hypothetical intervention to ensure that everyone takes the statistics test?

1. **Write the causal target parameter. What are the counterfactual outcomes?** Explain using notation and in words.

2. **Explain how to intervene on the SCM to get at the causal question/parameter of interest.** Refer back to `R` Lab 1 for this data structure's SCM.

3. **Implement the intervention described in step 2** by modifying the original data generating function created in `R` Lab 1. Call the new function `generate_data1_intervene`:

   (a) Copy and paste the data generating function (for this data structure) from `R` Lab 1, `generate_data1()`.

   (b) In addition to the argument in your function specifying the number of observations you want to generate (i.e., n), add arguments to your function based on the variables you want to intervene on. For example, if you want to intervene on A, add an argument `a` to your function:

```
> generate_data_example = function(n, a) {  # add the argument "a"
+
+   U.W = rnorm(n, mean = 1, sd = 1)
+   U.A = runif(n, min = 0, max = 1)
+   U.Y = rnorm(n, mean = 1, sd = 3)
+
+   W = U.W
+   A = as.numeric(W + U.A < 1.5)
+   Y = W + A + U.Y
+
+   O = data.frame(W, A, Y)
+
+   return(O)
+
+ }
```

   (c) Intervene on the endogenous variables of interest by setting them equal to the new arguments you added in the previous step. For example, if you want to intervene on the random variable "A" by setting it equal to the constant "a", then set `A` equal to `a` within your function:

```
> generate_data_intervene_example = function(n, a) {
+
+   U.W = rnorm(n, mean = 1, sd = 1)
+   U.A = runif(n, min = 0, max = 1)
+   U.Y = rnorm(n, mean = 1, sd = 3)
+
+   W = U.W
+   A = a # intervention is on A, so set A equal to a
+   Y = W + A + U.Y
+
+   X = data.frame(W, A, Y)
+
+   return(X)
+
+ }
```

4. **Evaluate** $\Psi^F(P_{U,X})$:

(a) Generate 100,000 observations of the data generating system you intervened on. Remember to add the values you want to intervene with as arguments. Store this in a dataframe. For example:

```
> X_example1 = generate_data_intervene_example(n = 100000, a = 1) # setting A = 1
```

(b) Extract the outcome(s) from the dataframe you generated in the previous step. These are your counterfactual outcomes. Remember that one way to extract a variable from a dataframe is using the dollar sign, $. For example:

```
> X_example1$Y # here we are extracting the counterfactual outcome Y1 from X_example1
```

(c) Evaluate $\Psi^F(P_{U,X})$ using the simulated counterfactual outcomes to obtain the true value of your target causal parameter. *Hint: take the mean of both counterfactual outcomes from the previous step, and subtract (in the right direction)!*

5. **Interpret $\Psi^F(P_{U,X})$.**

Solution:

1. Target causal parameter that would answer the causal question:

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

This is the difference in the counterfactual expected test score if all SPH students got 8 or more hours of sleep ($A = 1$) minus the counterfactual expected test score if all SPH students got less than 8 hours of sleep ($A = 0$), with no loss to follow-up (i.e., all students are forced to take the test).

The counterfactual $Y_{a,\Delta=1}$ is a random student's test score if, possibly contrary to fact, the student's sleep status had been $A = a$ and his/her test score was observed ($\Delta = 1$).

2. To get at $\Psi^F(P_{U,X})$, our causal parameter of interest, we would need to deterministically set $A = 1$ and $\Delta = 1$, then observe our outcome. Next, we would set $A = 0$ and $\Delta = 1$, then observe our outcome. Specifically, we would intervene on the structural equations for our endogenous variables as follows:

$$\begin{aligned}
W &= U_W \\
A &= a \in \{0,1\} \\
L &= W + a + U_L \\
\Delta &= 1 \\
Y &= L + 5^*a + 3^*W - 0.25^*a^*W + U_Y
\end{aligned}$$

Note that intervening on Δ does not change the true Y (note how Y is not a function of Δ). So, there's actually no need to intervene on Δ to generate counterfactual outcomes. The reason we do it comes in later, to get identifiability from the observed data.

```
> # 3. intervene on SCM/data generating function
> print(generate_data1_intervene)


function(n, a, delta) {

  # exogenous variables
  U.W = runif(n, min=0, max=1)
  U.A = runif(n, min=0, max=1)
  U.L = rnorm(n, mean=2, sd=1)
  U.Delta = runif(n, min=0, max=1)
  U.Y = rnorm(n, mean=72, sd=0.3)


  # endogenous variables
  W = U.W
  A = a # intervention on A
  L = W + A + U.L
  Delta = rep(delta, n) # intervention on Delta
  Y = L + 5*A + 3*W - 0.25*A*W + U.Y
  DeltaY = ifelse(Delta == 0, NA, Y)

  # store all variables in dataframe
  X = data.frame(W, A, L, Delta, Y, DeltaY)

  return(X)
}
<bytecode: 0x7ffd5d18c8c0>


> # 4. evaluate Psi.F
> X1_1 = generate_data1_intervene(n = 100000, a = 1, delta = 1)
> X1_0 = generate_data1_intervene(n = 100000, a = 0, delta = 1)


> # ATE defined in terms of true Y (under Delta = 1):
> Psi.F1 = mean(X1_1$Y) - mean(X1_0$Y)
> Psi.F1


[1] 5.878555
```

5. Here, $\Psi^F(P_{U,X}) = 5.88$. In words: the counterfactual expected test score would be 5.88 points higher if all students got 8 hours of sleep than if all students got less than 8 hours of sleep the night before their statistics test, with no loss to follow up.

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

Causal question 1: How would the expected exam score at the end of the study (i.e., after $t = 4$ days) have differed if all students got 8 or more hours of sleep every night during the entire study (i.e., at $t = 1, 2, 3, 4$ days) versus if all students got less than 8 hours of sleep every night during the entire study (i.e., at $t = 1, 2, 3, 4$ days)?

1. **Write the causal parameter that would answer Causal question 1. What are the counterfactual outcomes?** Explain using notation and in words.

2. **Explain how to intervene on the SCM to get at Causal question 1.** Refer back to `R` Lab 1 for this data structure's SCM.

3. **Implement the intervention you described in step 2** by modifying `generate_data2()` from the previous lab. Call the new function `generate_data2_intervene`.

   (a) Similar to the previous data structure, copy the function `generate_data2()` from `R` Lab 1 and add an argument to the function that takes in *a vector* of values you want to intervene with. Name that vector `abar`.

(b) Intervene on the endogenous variables of interest by setting them equal to the *position* of the vector you added in the previous step. Use brackets to subset by position the value of interest from the vector `abar`. Adding to the previous example, if you want to set $A(1) = a(1)$ and $A(2) = a(2)$:

```
> generate_data_intervene_example2 = function(n, abar) {
+
+   U.W = rnorm(n, mean = 1, sd = 1)
+   U.A1 = runif(n, min = 0, max = 1)
+   U.A2 = runif(n, min = 0, max = 1)
+   U.Y = rnorm(n, mean = 1, sd = 3)
+
+   W = U.W
+   A1 = abar[1] # subset first position of abar vector to set equal to a(1)
+   A2 = abar[2] # subset second position of abar vector to set equal to a(2)
+   Y = W + A1 + A2 + U.Y
+
+   X = data.frame(W, A1, A2, Y)
+
+   return(X)
+
+ }
```

4. **Evaluate** $\Psi^F(P_{U,X})$.

   (a) Generate 100,000 observations of the data generating system you intervened on. Remember to add the vector of values you want to intervene with, `abar` as an argument. Store this in a dataframe.

   ```
   > # setting A1 = 0 and A2 = 0
   > X_example2 = generate_data_intervene_example2(n = 100000, abar = c(0, 0))
   ```

   (b) Extract the simulated counterfactual outcomes from the dataframe (as was done in the previous data structure) and evaluate $\Psi^F(P_{U,X})$.

5. **Interpret.**

Solution:

Causal question 1

1. The target causal parameter that would answer the question:

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

This is the difference in expected counterfactual exam score if all students got 8 or more hours of sleep the previous 4 nights before their statistics test minus the expected counterfactual exam score if all students got less than 8 hours of sleep on the previous 4 nights before their statistics test.

2. The intervention on our SCM would look like this:

$$
\begin{aligned}
L(1) &= U_{L(1)} \\
A(1) &= a(1) \\
L(2) &= a(1) + L(1) + U_{L(2)} \\
A(2) &= a(2) \\
L(3) &= a(1) + L(1) + a(2) + L(2) + U_{L(3)} \\
A(3) &= a(3) \\
L(4) &= a(1) + L(1) + a(2) + L(2) + a(3) + L(3) + U_{L(4)} \\
A(4) &= a(4) \\
Y &= 0.3^{*}L(1) + a(1) + 0.5^{*}L(2) + a(2) + 0.7^{*}L(3) + a(3) + L(4) + a(4) - U_Y + 130
\end{aligned}
$$

With $\bar{a}(4) \in ((1,1,1,1),(0,0,0,0))$, to evaluate $\Psi^F(P_{U,X})$, we need to set $A(1) = A(2) = A(3) = A(4) = 1$, evaluate Y, then set $A(1) = A(2) = A(3) = A(4) = 0$ and evaluate Y.

```
> #3. intervene on SCM/data generating function
> print(generate_data2_intervene)
```

```
function(n, abar) {

  # exogenous variables
  U.L1 = rnorm(n, mean=0, sd=1)
  U.A1 = runif(n, min=0, max=1)
  U.L2 = rnorm(n, mean=0, sd=1)
  U.A2 = runif(n, min=0, max=1)
  U.L3 = rnorm(n, mean=0, sd=1)
  U.A3 = runif(n, min=0, max=1)
  U.L4 = rnorm(n, mean=0, sd=1)
  U.A4 = runif(n, min=0, max=1)
  U.Y = rnorm(n, mean=72, sd=3)

  # endogenous variables
  L1 = U.L1
  A1 = abar[1]
  L2 = A1 + L1 + U.L2
  A2 = abar[2]
  L3 = A1 + L1 + A2 + L2 + U.L3
  A3 = abar[3]
  L4 = A1 + L1 + A2 + L2 + A3 + L3 + U.L4
  A4 = abar[4]
  Y = 0.3*L1 + A1 + 0.5*L2 + A2 + 0.7*L3 + A3 + L4 + A4 - U.Y + 130

  O = data.frame(L1, A1, L2, A2, L3, A3, L4, A4, Y)

  return(O)
}
<bytecode: 0x7ffd582d1ca0>
```

```
> #4. evaluate Psi.F
> X2_1111 = generate_data2_intervene(n=100000, abar = c(1, 1, 1, 1))
> X2_0000 = generate_data2_intervene(n=100000, abar = c(0, 0, 0, 0))
> Psi.F2 = mean(X2_1111$Y) - mean(X2_0000$Y)
> Psi.F2

[1] 13.6673
```

5. $\Psi^F(P_{U,X}) = 13.67$. In other words, the counterfactual expected test score would be 13.67 points higher if all students got 8 hours of sleep for all 4 nights than if all students got less than 8 hours of sleep for all 4 nights before their statistics test.

Causal question 2: How does cumulative days getting 8 or more hours of sleep affect students' statistics exam scores at the end of the study? Specifically, say you are willing to assume a linear relationship between total number of days on which a student got 8 or more hours of sleep and expected exam score. How could you summarize how much the expected exam score would change per additional night on which a student got at least 8 hours of sleep?

1. **Write the causal parameter that answers Causal question 2.** *Hint: refer to 252E Lecture 1, slides titled: "Defining target parameters using a longitudinal marginal structural model."*

2. **Explain how to intervene on the SCM to answer Causal question 2.** Refer back to `R` Lab 1 for this data structure's SCM. *Hint: instead of only setting $\bar{A}(4) = 1$ or $\bar{A}(4) = 0$, list out all the possible ways we could intervene on this SCM (i.e., every possible $\bar{a}(4)$).*

3. **Implement the intervention.**
   Hint: Use the exact same function as Causal question 1, `generate_data2_intervene()`*, to intervene! Skip to the next question.*

4. **Evaluate $\Psi^F(P_{U,X})$.** For each possible sleep regime $\bar{a}(4)$, calculate the corresponding expected test score under that regime $E[Y_{\bar{a}}]$. Summarize these expected outcomes as a linear function of total number of nights with 8 or more hours of sleep.

   (a) Make a matrix of all 16 possible $\bar{a}(4)$ regimes, where each row is a single regime and column is $A(1), ..., A(4)$.

   Hint: Use the the `expand.grid()` function, which takes in a vector and creates a data frame with all the possible permutations of the elements in that vector. Use the `colnames()` function to name the columns of matrix of $\bar{a}(4)$ permutations:

   ```
   > # matrix of every possible abar permutation
   > abar_mat = expand.grid(c(0,1), c(0,1), c(0,1), c(0,1))
   > # make column names each intervention node
   > colnames(abar_mat) = c("A1", "A2", "A3", "A4")
   ```

   (b) Create two new vectors of length 16 filled with `NA`s called `sum.abar` and `EY.abar`. In the next step, we will populate `sum.abar` with each regime's cumulative treatment and `EY.abar` with each regime's expected counterfactual outcome.

   (c) Create a `for` loop to compute the expected counterfactual outcome for each treatment regime. Recall that this is the syntax to create a for loop from `i` in `1:n`:

   ```
   > for (i in 1:n) {
   +   # insert code within for loop here
   + }
   ```

   Within a for loop from `i` in `1:16`, do the following:

      i. Generate a new data frame in which $\bar{A}(4)$ is intervened on using the i^{th} treatment regime in the `abar_mat` matrix.

Hint: Use the `generate_data2_intervene()` function to generate the new, intervened-on data, with the `abar` argument equal to the i^{th} row of `abar_mat`. Set this new data equal to `X`:

```
> X = generate_data2_intervene(n = 100000, abar = as.numeric(abar_mat[i,]))
```

ii. Get the cumulative treatment for the i^{th} regime, and save it to the i^{th} row of `sum.abar`.

Hint: Use the `rowSums` function on the i^{th} regime in `abar_mat`. Save this cumulative treatment to the i^{th} position of the `sum.abar` vector:

```
> sum.abar[i] = rowSums(abar_mat)[i]
```

iii. Get the mean counterfactual outcome under i^{th} regime, and save it in the `i`th position of `EY.abar`:

```
> EY.abar[i] = mean(X$Y)
```

Pause here: what did we just do?

- For each of the 16 possible treatment regimes $\bar{a}(4)$ (stored in `abar_mat`), we got the expected counterfactual outcome $E[Y_{\bar{a}}]$ (stored in `EY.abar`) under that regime.
- We also have a corresponding summary measure (in this case, a simple sum) of each treatment regime ($\sum \bar{a}(4)$, stored in `sum.abar`).

5. **Evaluate** $\Psi^F(P_{U,X})$**.** Recall that we are assuming that the expected outcome under each treatment regime $E[Y_{\bar{a}}]$ varies as a linear function of cumulative treatment $\sum \bar{a}(4)$. Use the `glm()` function to obtain the coefficients of this linear fit.

6. **Interpret** $\Psi^F(P_{U,X})$. Recall that our causal question of interest is: "how much does the expected exam score change per additional night on which a student got at least 8 hours of sleep?" Which coefficient from the previous step answers this question?

7. **Bonus:** Is the linear MSM you wrote down correctly specified? Why or why not? If you are not willing to assume that your MSM is correctly specified, but you are still interested in a linear summary of how the expected counterfactual exam score varies as a function of cumulative number of nights on which a student got more than 8 hours of sleep, how would you modify your target parameter?

8. **Extra bonus!** Plot the true underlying values $E[Y_{\bar{a}}]$ for each $\bar{a}$ and their projection onto the linear working model.

Solution: Causal question 2

1. We will use a Marginal Structural Model (MSM) to help us answer this question:

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

This MSM lets us summarize how the expectation of counterfactual statistics test scores varies as a function of cumulative sleep. In particular, we are interested in β_1 as it will tell us how the above summary measure (total number of full nights of sleep) affects the outcome (i.e., for one additional night of 8 or more hours of sleep, what is the change in students' mean counterfactual test score?).

As written above, we are assuming a correctly specified MSM. In practice, we generally can't assume that the MSM is correctly specified, so we use a working MSM, and instead, the above parameter, β, gives us the projection of the true "causal curve" $E_{U,X}[Y_{\bar{a}}]$, for all possible $\bar{a}$ onto a linear summary model, $m(\bar{a}|\beta)$ (see the bonus question).

2. We would use the same intervention on our SCM; however, instead of setting $\bar{A}(4) = 1$, then $\bar{A}(4) = 0$, we would need to intervene to generate every possible $\bar{a}(4)$. That is:

$$
\begin{gathered}
a(1) = 0, a(2) = 0, a(3) = 0, a(4) = 0 \\
a(1) = 1, a(2) = 0, a(3) = 0, a(4) = 0 \\
a(1) = 0, a(2) = 1, a(3) = 0, a(4) = 0 \\
\ldots \\
a(1) = 1, a(2) = 1, a(3) = 1, a(4) = 1
\end{gathered}
$$

There are $2^4 = 16$ possible permutations of $\bar{a}(4)$.

3. We will use the same function as Causal question 1 (above) to intervene.

```
> #4. evaluate Psi.F
> # matrix of every possible txt regime permutation
> # Each column is an intervention at time t, or A(t).
> # Each row is a possible instance of abar.
> abar_mat = expand.grid(c(0,1), c(0,1), c(0,1), c(0,1))
> # name the columns
> colnames(abar_mat) = c("A1", "A2", "A3", "A4")

> # create an empty vector of NAs for the cumulative abars and mean counterfactual outcomes
> sum.abar = EY.abar = rep(NA, 16)

> # "For" loop to generate counterfactual outcomes.
> # For every rows of the dataframe of possible a(t)'s, apply the function
> # generate_data2_intervene() to generate data that has
> # been intervened on with that abar regime.
> # Take the mean of the Y's of interest.
> for(i in 1:16) {
+
+   X = generate_data2_intervene(100000, abar = as.numeric(abar_mat[i,]))
+   sum.abar[i] = rowSums(abar_mat)[i]
+   EY.abar[i] = mean(X$Y)
+
+ }

> # use ordinary least squares regression to obtain beta coefficients.
> # the target parameter is the coefficient value that minimize the L2 risk function
> MSM = glm(EY.abar ~ sum.abar)
> MSM

Call:  glm(formula = EY.abar ~ sum.abar)

Coefficients:
(Intercept)     sum.abar
     57.995        3.407

Degrees of Freedom: 15 Total (i.e. Null);  14 Residual
Null Deviance:            266
Residual Deviance: 80.28         AIC: 77.21
```

```
> # beta 1 coefficient
> TrueMSMbeta1 = MSM$coefficients[[2]]
> TrueMSMbeta1

[1] 3.406645
```

5. The true dose-response curve's intercept is 57.995 and slope is 3.407. This means that, if we believe this MSM is correct, cumulative days with 8 more hours of sleep has a positive linear effect on the mean counterfactual statistics test scores. For one more night of 8 or more hours of sleep, students' statistics test scores increase by 3.407 points, on average.

Bonus: The linear MSM we wrote down is not correctly specified. To see why not, note that even though the A at each time point has the same direct effect on Y, an intervention on $A(1)$ will have a different effect on Y than an intervention on $A(4)$, because an intervention on $A(1)$ will also affect Y via its effects on intervening L's.

We could modify our target parameter by defining a *working* MSM, in which the target parameter β is defined as a projection of the true underlying causal curve $E_{U,X}(Y_a)$ onto a linear working model $m(a|\beta)$:

$$\beta(P_{U,X}|m) = argmin_{\beta}\ E_{U,X}\left[\sum_{\bar{a}\in\mathcal{A}} \left(Y_{\bar{a}} - m(\bar{a}|\beta)\right)^2\right]$$

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

Thus, projecting the causal dose-response curve onto the working MSM yields coefficients of $\beta_0 = 57.995$ and $\beta_1 = 3.407$. In other words, summarizing the true causal dose-response curve with a linear working MSM suggests that cumulative sleep has a positive linear effect on mean statistics exam score.

Extra: MSM with different projection function

We can modify the working MSM further by using a projection function, $g_n(\bar{A}(K))$ (i.e., the marginal probability of receiving each treatment regime of interest). Adding stabilized weights allows for weaker positivity assumptions – the target parameter is still defined if some regimes don't occur. When we stabilize, however, we are working with a different causal parameter than without stabilization.

To do this, we will need to generate a large number of observations from data generating system 2 *without* intervening on the system, and calculate the marginal probability of each of the 16 observed treatment regimes. We will repeat our regression of $E[Y_{\bar{a}}]$ for each possible $\bar{a}$ on the linear MSM, but now we will fit a weighted linear regression, where each $E[Y_{\bar{a}}]$ gets a weight corresponding to its probability of occuring in the observed data.

```
> # generate a large number of observations
> ObsData2 = generate_data2(n=100000)


> # initialize a vector for the marginal probability of each permutation of abar(4)
> # in other words, g(abar(4))
> g.abar = rep(NA, 16)


> # calculate the marginal probability of each of the 16 observed treatment regimes
> for(i in 1:16){
```

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


> # run the regression E[Yabar] on sum of abar, with new weights
> MSM_wts = glm(EY.abar ~ sum.abar, weights = g.abar)
> MSM_wts


Call:  glm(formula = EY.abar ~ sum.abar, weights = g.abar)

Coefficients:
(Intercept)     sum.abar
     57.982        3.406

Degrees of Freedom: 15 Total (i.e. Null);  14 Residual
Null Deviance:          16.66
Residual Deviance: 5.001        AIC: 77.16


> # beta1 coefficient
> TrueMSMbeta1_wts = MSM_wts$coefficients[[2]]
> TrueMSMbeta1_wts


[1] 3.405607
```

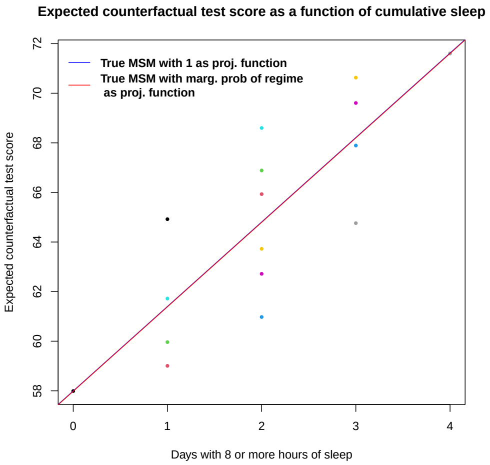
Now, the true dose-response curve's slope is 3.406. This means that, if we believe this MSM is correct, cumulative days with 8 more hours of sleep has a positive linear effect on the mean counterfactual statistics test scores. For one more night of 8 or more hours of sleep, students' statistics test scores increase by 3.406 points, on average.

Extra bonus!

```
> # plot the expected test score as a function of cumulative sleep (unweighted)
> plot(sum.abar, EY.abar,
+      main='Expected counterfactual test score as a function of cumulative sleep',
+      xlab='Days with 8 or more hours of sleep',
+      ylab='Expected counterfactual test score',
+      pch = 16,
+      col=as.factor(EY.abar),
+      cex = .7)


> # add the true value of the causal parameter beta to the graph (in blue).
> abline(MSM, col='blue')


> # add the true causal curve (weighted) to the graph (in red).
> abline(MSM_wts, col = "red")
> # add legend
> legend("topleft",
+        legend = c("True MSM with 1 as proj. function", "True MSM with marg. prob of regime \n as proj
+        col = c("blue", "red"),
+        lty = 1,
+        text.font=2,
+        box.lty=0)
```



Solution Fig. 1: Plot the true values of expected counterfactual outcomes $E[Y_{\bar{a}}]$ as a function of $\sum_{t=1}^{4} a(t)$ and their corresponding projection onto a linear MSM. Each of the 16 colored dots represents the expected counterfactual test score of each of the 16 treatment regimes. Both lines are true MSMs; the blue line is the true MSM without weights, and the red line is the true MSM with stabilizing weights. As we can see, the two lines are very similar, meaning that there is little variability in the marginal probabilities for the regimes. Put another way, all of the regimes occur at roughly similar frequencies.

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

Causal question: How would the counterfactual probability of becoming sick differ by the time of the test under an intervention to get 8 or more hours of sleep for 2 nights before a statistics test versus an intervention to get less than 8 hours of sleep for 2 nights before a statistics test?

1. **Write the causal target parameter that would answer the causal question posed for data structure 3. What are the counterfactual outcomes?** Explain using notation and in words.
2. **Explain how to intervene on the SCM to get at the causal question/parameter of interest.** Refer back to `R` Lab 1 for each SCM.
3. **Implement the intervention described in step 2** by using `generate_data3()` from the previous lab. Call this function `generate_data3_intervene()`.
4. **Evaluate the causal parameter via simulated counterfactuals.**
5. **Interpret the estimand you generated in the previous step.**

Solution:

1. Our target causal parameter here is:

$$
\begin{aligned}
\Psi^F(P_{U,X}) =& E[Y(3)_{\bar{a}(2)=1}] - E[Y(3)_{\bar{a}(2)=0}] \\
=& P\big(Y(3)_{\bar{a}(2)=1} = 1\big) - P\big(Y(3)_{\bar{a}(2)=0} = 1\big)
\end{aligned}
$$

This is the difference in counterfactual probability of getting sick if all SPH students got 8 or more hours of sleep for both nights ($\bar{A}(2) = 1$) minus the counterfactual probability of getting sick if all SPH students got fewer than 8 hours of sleep for both nights ($\bar{A}(2) = 0$). The counterfactual $Y(3)_{\bar{a}(2)}$ is the student's illness status on day 3 of the study if, possibly contrary to fact, the student's sleep status was $\bar{A}(2) = \bar{a}(2)$.

2. We would intervene on our SCM by setting $A(t) = a(t)$ in the following way to answer our causal question:

$$
\begin{aligned}
&L(1) = U_{L(1)} \\
&A(1) = a(1) \\
&Y(2) = \mathbb{I}\big[U_{Y(2)} < \mathit{expit}(L(1) - 2^{*}a(1) - 6)\big] \\
&L(2) = \begin{cases} a(1) + L(1) + U_{L(2)} & \text{if } Y(2) = 0 \\ \texttt{NA} & \text{if } Y(2) = 1 \end{cases} \\
&A(2) = a(2) \\
&Y(3) = \begin{cases} \mathbb{I}\big[U_{Y(3)} < \mathit{expit}(L(1) - 2^{*}a(1) + L(2) - a(2))\big] & \text{if } Y(2) = 0 \\ 1 & \text{if } Y(2) = 1 \end{cases}
\end{aligned}
$$

Note that while we can define an intervention after an individual has been seen to get sick, such an intervention will have no effect on the final outcome if an individual has already gotten sick. In other words, intervening to set sleep status on the second night ($A(2) = a(2)$) can only affect whether a student becomes sick by the time of the test ($Y(3)$) if the student has not already become sick before the second night (i.e., if $Y(2) = 0$).

Solution:

```
> #3. intervene on SCM/data generating function
> print(generate_data3_intervene)


function(n, abar) {

  # exogenous variables
  U.L1 = rnorm(n, mean=0, sd=1)
  U.A1 = runif(n, min=0, max=1)
  U.Y2 = runif(n, min=0, max=1)
  U.L2 = rnorm(n, mean=0, sd=1)
  U.A2 = runif(n, min=0, max=1)
  U.Y3 = runif(n, min=0, max=1)
  # endogenous variables
  L1 = U.L1
  A1 = abar[1]
  Y2 = as.numeric(U.Y2 < plogis(L1 - 2*A1 - 6))
  L2 = ifelse(Y2 == 1, NA, A1 + L1 + U.L2)
  A2 = abar[2]
  Y3 = ifelse(Y2 == 1, 1, as.numeric(U.Y3 < plogis(L1 - 2*A1 + L2 - A2)))

  X = data.frame(L1, A1, Y2, L2, A2, Y3)

  return(X)

}
<bytecode: 0x7ffd5806fef8>


> #4. evaluate Psi.F
> X3_11 = generate_data3_intervene(n = 100000, abar = c(1, 1))
> X3_00 = generate_data3_intervene(n = 100000, abar = c(0, 0))
> Psi.F3 = mean(X3_11$Y3) - mean(X3_00$Y3)
> Psi.F3


[1] -0.25933
```

5. $\Psi^F(P_{U,X}) =$ -0.26. The counterfactual probability of getting sick would be -26% higher if all students got 8 or more hours of sleep for 2 nights before their statistics test than if all students got less than 8 hours of sleep for 2 nights before their statistics test.

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

Causal question: How would the counterfactual probability of becoming sick differ under an intervention to get 8 or more hours of sleep for 2 nights before a statistics test versus an intervention to get less than 8 hours of sleep for 2 nights before a statistics test, forcing all students to stay in the class for the time of observation?

1. **Write the causal target parameter that would answer the causal question posed for data structure 4. What are the counterfactual outcomes?** Explain using notation and in words.
2. **Explain how to intervene on the SCM to get at the causal question/parameter of interest.** Refer back to `R` Lab 1 for this data structure's SCM.
3. **Implement the intervention described in step 2** by using `generate_data4()` from the previous lab. Call this function `generate_data4_intervene()`.
4. **Evaluate the causal parameter via simulated counterfactuals.**
5. **Interpret the estimand you generated in the previous step.**

Solution:

1. Target causal parameter:

$$\begin{aligned}\Psi^F(P_{U,X}) =& E[Y(3)_{\bar{a}(2)=1,\bar{c}(2)=0}] - E[Y(3)_{\bar{a}(2)=0,\bar{c}(2)=0}]\\ =& P\big(Y(3)_{\bar{a}(2)=1,\bar{c}(2)=0} = 1\big) - P\big(Y(3)_{\bar{a}(2)=0,\bar{c}(2)=0} = 1\big)\end{aligned}$$

This is the difference in counterfactual probability of getting sick if all SPH students got 8 or more hours of sleep every night (setting $\bar{A}(2) = 1$) minus the counterfactual probability of getting sick if all SPH students got fewer than 8 hours of sleep every night (setting $\bar{A}(2) = 0$), ensuring no loss to follow-up by setting $\bar{C}(2) = 0$ (i.e., intervening to ensure that no students drop out of the class). The counterfactual $Y(3)_{\bar{a}(2),\bar{c}(2)=0}$ is the student's illness status on day 3 of the study if, possibly contrary to fact, the student's sleep status was $\bar{A}(2) = \bar{a}(2)$ and he/she remained in the class $\bar{C}(2) = 0$.

2. Intervene on the SCM by setting $\bar{A}(t) = \bar{a}(t)$ and $\bar{C}(t) = \bar{c}(t)$ as follows:

$$\begin{aligned}
&L(1) = U_{L(1)}\\
&C(1) = 0\\
&A(1) = a(1)\\
&Y(2) = \begin{cases}\mathbb{I}\big[U_{Y(2)} < \textit{expit}(L(1) - 2{*}a(1) - 6)\big] & \text{if } C(1) = 0\\ \texttt{NA} & \text{if } C(1) = 1\end{cases}\\
&L(2) = \begin{cases}a(1) + L(1) + U_{L(2)} & \text{if } Y(2) = 0 \text{ and } C(1) = 0\\ \texttt{NA} & \text{if } Y(2) = 1 \text{ or } C(1) = 1\end{cases}\\
&C(2) = 0\\
&A(2) = a(2)\\
&Y(3) = \begin{cases}\mathbb{I}\big[U_{Y(3)} < \textit{expit}(L(1) - 2{*}a(1) + L(2) - a(2))\big] & \text{if } Y(2) = 0 \text{ and } C(2) = 0\\ 1 & \text{if } Y(2) = 1\\ \texttt{NA} & \text{if } Y(2) = 0 \text{ and } C(2) = 1\end{cases}
\end{aligned}$$

Similar to the above example, interventions on both sleep and to prevent loss to follow-up that occur after an individual has gotten sick will have no effect on the final counterfactual outcome, an indicator of becoming sick by the time of the test. That is, once you have become sick and we have had a change to observe it, subsequently dropping out of the class will not change this.

```
> #3. intervene on SCM/data generating function
> print(generate_data4_intervene)


function(n, abar, cbar) {

  # exogenous variables
  U.L1 = rnorm(n, mean=0, sd=1)
  U.C1 = runif(n, min=0, max=1)
  U.A1 = runif(n, min=0, max=1)
  U.Y2 = runif(n, min=0, max=1)
  U.L2 = rnorm(n, mean=0, sd=1)
  U.C2 = runif(n, min=0, max=1)
  U.A2 = runif(n, min=0, max=1)
  U.Y3 = runif(n, min=0, max=1)

  # endogenous variables
  L1 = U.L1
  C1 = rep(cbar[1], n) # intervention on C1
  A1 = abar[1] # intervention on A1
  Y2 = as.numeric(U.Y2 < plogis(L1 - 2*A1 - 6))
  L2 = ifelse(Y2 == 1, NA, A1 + L1 + U.L2)
  C2 = rep(cbar[2], n) # intervention on C2
  A2 = abar[2] # intervention on A2
  Y3 = ifelse(Y2 == 1, 1, as.numeric(U.Y3 < plogis(L1 - 2*A1 + L2 - A2)))

  X = data.frame(L1, C1, A1, Y2, L2, C2, A2, Y3)

  return(X)

}
<bytecode: 0x7ffd57f384d8>


> #4. evaluate Psi.F
> X4_11 = generate_data4_intervene(n = 100000, abar = c(1, 1), cbar = c(0, 0))
> X4_00 = generate_data4_intervene(n = 100000, abar = c(0, 0), cbar = c(0, 0))
> Psi.F4 = mean(X4_11$Y3) - mean(X4_00$Y3)
> Psi.F4


[1] -0.26174
```

5. $\Psi^F(P_{U,X})$ = -0.26. The counterfactual probability of getting sick would be 26% lower if all students got 8 or more hours of sleep for 2 nights before their statistics test than if all students got less than 8 hours of sleep for 2 nights before their statistics test, forcing all students to take the exam.

Notice that this answer is almost equal to the parameter evaluated in the previous data structure. This is because in both scenarios all students are forced to take the test – in the first data structure we don't allow for censoring, and in this data structure we force everyone to be uncensored. Any difference is due to random number generation of the exogenous variables. Can you think of a scenario where intervention on the censoring might change the true value of the target parameter?

Take home message: go to sleep as soon as you finish this lab!

3 For Your Project: Evaluating Target Causal Parameters

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

1. **Defining your causal question**
   - (a) What is the causal question (or questions) of interest for your dataset?
   - (b) What is the ideal experiment that would answer your causal question?
   - (c) Which of your variables variables would you intervene on to answer your causal question(s)? What values would you set them equal to?
   - (d) What outcomes are you interested in? Measured when?
2. **Target parameter and counterfactual outcomes**
   - (a) What are your counterfactual outcomes, and how would you explain them in words?
   - (b) Come up with a target parameter that would answer your causal question.
     - What aspects of the counterfactual outcome distribution are you interested in contrasting?
     - What contrast are you interested in (e.g., absolute difference? relative difference? MSMs? conditional on subgroups?)?
3. **Intervention on SCM**
   - (a) How would you intervene on the SCM you came up with to evaluate the causal target parameter?
   - (b) Implement this intervention computationally.
4. **Evaluate $\Psi^F(P_{U,X})$**
   - (a) Using simulations, generate many counterfactual outcomes.
   - (b) Evaluate $\Psi^F(P_{U,X})$.
   - (c) Write a sentence interpreting your $\Psi^F(P_{U,X})$.

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?