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

Advanced Topics in Causal Inference

Assigned: September 22, 2020
Lab due: September 29, 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:
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 $\Psi^F(P_{U,X})$.**

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.**

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.

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.**

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.**

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?