

R Lab 3 - Understanding Time Dependent Confounding and Identifiability in Longitudinal Context

Advanced Topics in Causal Inference

Assigned: September 29, 2020

Lab due: October 6, 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:

Translate causal questions into target causal parameters and intervene on the data generating processes described by Structural Causal Models (SCMs) to evaluate the true value of these parameters.

This lab:

1. Review the concept of identifiability.
2. Determine which assumptions let us achieve identifiability.
3. Write the target causal parameter (a function of the counterfactual or “post intervention” distribution) as a function of the observed data distribution.
4. Obtain the value of the statistical estimand.
5. Understand the challenges posed by time-dependent confounding.

Next lab:

Implementation of longitudinal IPTW to estimate the intervention specific mean and the parameters of a marginal structural model.

1 Introduction and Motivation

In previous labs we examined the counterfactual distributions (and parameters of these distributions) of the variables that help us answer our causal questions of interest.

In reality, we won’t know the counterfactual distribution (e.g., we can’t observe exogenous errors or counterfactual outcomes, and we don’t know the true data generating process). Instead, we see instances of data generated from the *observed* data distribution.

So, we now would like to be able to write our target causal parameter as a parameter of our observed data distribution P_0 . If we can do this, we have achieved *identifiability*. Having defined a target statistical parameter (or estimand), we can estimate it using our observed data (typically in this class, n i.i.d. copies of the observed random variable O ; coming up in the next lab).



Fig. 1: Identification.

2 This lab

Last time, you wrote down target causal parameters inspired by the causal questions posed to you by your GSR. However, in order to actually evaluate effects of sleep on health and performance outcomes, you'll need to come up with a function of the distribution of the observed data that your professor hands to you.

Since you are dealing with studies in which you are interested in the effects of multiple interventions (e.g., cumulative sleep over time), to achieve identifiability, you will need to consider assumptions beyond ones you've learned in the point-treatment setting.

The first data structure works off of new variable definitions and data generating processes. The last three data structures build off of previous labs: refer back to R Lab 1 for variable definitions and specific data generating processes for data structures 1, 2, and 4.

2.1 To turn in:

Note: there are separate questions for Data Structure 0

For Data Structures 1, 2, and 4, answer the following questions:

1. **Is the true $P_{U,X}$ an element of (i.e., described by or compatible with) the SCM presented?**
Refer back to R Lab 1 for the true $P_{U,X}$.
2. We will present a target causal parameter. **Is the target causal parameter (a parameter of $P_{U,X}$) identified (as a parameter of P_0) under the standard, point treatment randomization assumption/back door criteria? Why or why not?**
3. **If the target parameter is not identified in the previous question, what are the alternative assumptions under which the parameter would be identified?**
4. **What is the corresponding statistical estimand, $\Psi(P_0)$, under these assumptions?**

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

Recall that in Causal I, we learned a common identification result for the expectation of the counterfactual outcome under a *point treatment intervention* (such as the average treatment effect). Specifically, for $O = (W, A, Y)$, under the randomization assumption ($Y_a \perp A|W$), or if W satisfied the back door criteria with respect to the effect of A on Y , (together with the positivity assumption) we had the following result:

$$E[Y_a] = \sum_w E_0[Y|A = a, W = w] \times P_0(W = w)$$

The right hand side of the equation is called the *point-treatment g-computation formula*.

In the following example we use a simple discrete data structure to illustrate how this identification result can break down in the *longitudinal* setting – that is, when we are interested in evaluating the effects of interventions on more than one variable.

SCM:

$U = (U_{\bar{L}(2)}, U_{\bar{A}(2)}, U_Y) \sim P_U$, where we assume all U s are independent.

$X = (L(1), A(1), L(2), A(2), Y)$

and f_X is:

$$\begin{aligned} L(1) &= f_{L(1)}(U_{L(1)}) \\ A(1) &= f_{A(1)}(U_{A(1)}, L(1)) \\ L(2) &= f_{L(2)}(U_{L(2)}, L(1), A(1)) \\ A(2) &= f_{A(2)}(U_{A(2)}, \bar{L}(2)) \\ Y &= f_Y(U_Y, \bar{L}(2), \bar{A}(2)) \end{aligned}$$

$P_{U,X}$ (implying one particular joint distribution of (U, X) , in other words, is an element of the above SCM):

Exogenous variables:

$U_{L(t)}$ for $t = 1, 2 \sim \text{Uniform}(\min = 0, \max = 1)$

$U_{A(t)}$ for $t = 1, 2 \sim \text{Uniform}(\min = 0, \max = 1)$

$U_Y \sim \text{Uniform}(\min = 0, \max = 1)$

Structural equations F and endogenous variables:

$L(1) = \mathbb{I}[U_{L(1)} < 0.5]$

$A(1) = \mathbb{I}[U_{A(1)} < \text{expit}(0.3 - L(1))]$

$L(2) = \mathbb{I}[U_{L(2)} < \text{expit}(-2 + 1.8 * A(1) + 2 * L(1))]$

$A(2) = \mathbb{I}[U_{A(2)} < \text{expit}(L(2) + L(1))]$

$Y = \mathbb{I}[U_Y < \text{expit}(-3 + 1.3 * A(1) + 1.7 * A(2) + 1.3 * L(1) + 1.7 * L(2))]$

And the true value of the target causal parameter of interest is (optional: verify this yourself!):

$$\Psi^F(P_{U,X}) = E_{U,X}[Y_{\bar{a}=1}] = 0.7922$$

Specifically, below we'll show that the point treatment g-computation formula (which conditions on either $L(1)$ and $L(2)$, or on $L(1)$ only) does not equal the true value of our longitudinal target causal parameter:

1. **Generate a large amount of data according to $P_{U,X}$, above.** In R, generate a large number of observations of O (say, $n = 1,000,000$) according to the above data-generating process.
2. **If we use the *point-treatment* g-computation formula and condition on $L(1)$ and $L(2)$, what is the statistical estimand $\Psi^{(1)}(P_0)$ equal to? Is it equal to the true value of the target causal parameter?**

$$\Psi^{(1)}(P_0) = \sum_{\bar{l}(2)} E_0[Y|\bar{A}(2) = 1, \bar{L}(2) = \bar{l}(2)] \times P_0(\bar{L}(2) = \bar{l}(2)) = ???$$

- (a) Notice that the sum of $\Psi^{(1)}(P_0)$ is over every permutation of $\bar{L}(2) = \bar{l}(2)$:

Permutation 1 = ($L(1) = 1, L(2) = 1$)

Permutation 2 = ($L(1) = 1, L(2) = 0$)

Permutation 3 = ($L(1) = 0, L(2) = 1$)

Permutation 4 = ($L(1) = 0, L(2) = 0$)

For each of these permutations we need to compute:

$$E_0[Y|\bar{A}(2) = 1, \bar{L}(2) = \bar{l}(2)] \times P_0(\bar{L}(2) = \bar{l}(2))$$

We'll demonstrate how to do this for Permutation 1. Then, you can repeat this derivation for all other permutations, using Permutation 1 as a model. Let's break this quantity under Permutation 1 in two parts:

$$\underbrace{E_0[Y|\bar{A}(2) = 1, \bar{L}(2) = 1]}_{\text{Part 1}} \times \underbrace{P_0(\bar{L}(2) = 1)}_{\text{Part 2}}$$

Part 1 Compute $E_0[Y|\bar{A}(2) = 1, \bar{L}(2) = 1]$ by subsetting the values of Y for which $A1$, $A2$, $L1$, and $L2$ are 1, and take the mean:

```
> mean(Y[A1 == 1 & A2 == 1 & L1 == 1 & L2 == 1])
```

Part 2 Compute $P_0(\bar{L}(2) = 1)$ by obtaining the proportion of times where $L1$ and $L2$ are 1:

```
> mean(L1 == 1 & L2 == 1)
```

- (b) Multiply **Part 1** and **Part 2** together.
- (c) Repeat this process for all other permutations of $\bar{L}(2)$. Then, sum all of these quantities together to obtain the statistical estimand, $\Psi^{(1)}(P_0)$.
- (d) Is $\Psi^{(1)}(P_0)$ equal to $\Psi^F(P_{U,X})$? Have we achieved identifiability?
3. **If we use the *point-treatment* g-computation formula and condition on $L(1)$, what is the statistical estimand $\Psi^{(2)}(P_0)$ equal to? Is it equal to the true value of the target causal parameter?** Solve $\Psi^{(2)}(P_0)$ computationally; in other words, use R to evaluate $\Psi^{(2)}(P_0)$, as in the previous problem. *Hint: be sure to sum over the correct permutations!*

$$\Psi^{(2)}(P_0) = \sum_{l(1)} E_0[Y|\bar{A}(2) = 1, L(1) = l(1)] \times P_0(L(1) = l(1)) = ???$$

4. **If we use the *longitudinal* g-computation formula, what is the statistical estimand $\Psi^{(3)}(P_0)$ equal to? Is it equal to the true value of the target causal parameter?**

$$\Psi^{(3)}(P_0) = \sum_{\bar{l}(2)} E_0[Y|\bar{A}(2) = 1, \bar{L}(2) = \bar{l}(2)] \times P_0(L(2) = \bar{l}(2) | A(1) = 1, L(1) = \bar{l}(1)) \times P_0(L(1) = \bar{l}(1)) = ???$$

- (a) **Solve** $\Psi^{(3)}(P_0)$ **computationally**. In other words, use R to evaluate $\Psi^{(3)}(P_0)$ (as in the previous two problems).
- (b) **Bonus!** Solve $\Psi^{(3)}(P_0)$ analytically. Again, notice that the sum of $\Psi^{(3)}(P_0)$ is over every permutation of $\bar{L}(2) = \bar{l}(2)$. We'll demonstrate how to get started for $(L(1) = 1, L(2) = 1)$. Then, you can repeat this derivation for all other permutations of $\bar{L}(2)$. First, let's break $\Psi^{(3)}(P_0)$ into three parts:

$$\Psi^{(3)}(P_0) = \sum_{\bar{l}(2)} \underbrace{E_0[Y|\bar{A}(2) = 1, \bar{L}(2) = \bar{l}(2)]}_{\text{Part 1}} \times \underbrace{P_0(L(2) = l(2)|A(1) = 1, L(1) = l(1))}_{\text{Part 2}} \times \underbrace{P_0(L(1) = l(1))}_{\text{Part 3}}$$

Part 1 Compute $E_0[Y|\bar{A}(2) = 1, \bar{L}(2) = 1]$:

$$E[Y|\bar{A}(2) = 1, \bar{L}(2) = 1] = E[\mathbb{I}[U_Y < \text{expit}(-3 + 1.3^*A(1) + 1.7^*A(2) + 1.3^*L(1) + 1.7^*L(2))|\bar{A}(2) = 1, \bar{L}(2) = 1]]$$

The expectation of an indicator is just the probability that indicator equals 1:

$$= P\left(U_Y < \text{expit}(-3 + 1.3^*A(1) + 1.7^*A(2) + 1.3^*L(1) + 1.7^*L(2))|\bar{A}(2) = 1, \bar{L}(2) = 1\right)$$

Substitute in the constant values we're conditioning on:

$$= P\left(U_Y < \text{expit}(-3 + 1.3^*1 + 1.7^*1 + 1.3^*1 + 1.7^*1)\right)$$

The probability a uniform random variable between 0 and 1 is less than a constant value is just that constant value:

$$= \text{expit}(-3 + 1.3^*1 + 1.7^*1 + 1.3^*1 + 1.7^*1)$$

Part 2 Compute $P_0(L(2) = 1|A(1) = 1, L(1) = 1)$:

$$\begin{aligned} P(L(2) = 1|A(1) = 1, L(1) = 1) &= P\left(\mathbb{I}[U_{L(2)} < \text{expit}(-2 + 1.8^*A(1) + 2^*L(1))|A(1) = 1, L(1) = 1] = 1\right) \\ &= P\left(\mathbb{I}[U_{L(2)} < \text{expit}(-2 + 1.8^*1 + 2^*1)] = 1\right) \\ &= \text{expit}(-2 + 1.8^*1 + 2^*1) \end{aligned}$$

Part 3 Compute $P_0(L(1) = 1)$:

$$P(L(1) = 1) = P(\mathbb{I}[U_{L(1)} < 0.5] = 1) = 0.5$$

Multiply **Part 1**, **Part 2**, and **Part 3** together.

Repeat this process for all other permutations of $\bar{L}(2)$. Then, sum over all of these quantities together to obtain the statistical estimand, $\Psi^{(3)}(P_0)$.

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

$U = (U_W, U_A, U_L, U_\Delta, U_Y) \sim P_U$. Assume U s are jointly independent.
Structural equations, F :

$$\begin{aligned} W &= f_W(U_W) \\ A &= f_A(W, U_A) \\ L &= f_L(W, A, U_L) \\ \Delta &= f_\Delta(W, A, L, U_\Delta) \\ Y &= f_Y(W, A, L, U_Y) \end{aligned}$$

Is the true $P_{U,X}$ described by the SCM presented? Refer back to the first data structure of R Lab 1 for the true $P_{U,X}$.

2. Target causal parameter:

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

Is this target causal parameter (a parameter of $P_{U,X}$) identified (as a parameter of P_0) under the standard, point treatment randomization assumption/back door criteria? Why or why not?

3. **If the target parameter is not identified in the previous question, what are the alternative assumptions under which the parameter would be identified?**
4. **What is the corresponding statistical estimand, $\Psi(P_0)$, under these assumptions?**

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

$U = (U_{L(t)}, U_{A(t)}, U_Y), t = 1, \dots, 4 \sim P_U$. Assume U s are jointly independent.
Structural equations, F :

$$\begin{aligned} L(1) &= f_{L(1)}(U_{L(1)}) \\ A(1) &= f_{A(1)}(L(1), U_{A(1)}) \\ L(t) &= f_{L(t)}(\bar{L}(t-1), \bar{A}(t-1), U_{L(t)}) \text{ for } t = 2, \dots, 4 \\ A(t) &= f_{A(t)}(\bar{A}(t-1), \bar{L}(t), U_{A(t)}) \text{ for } t = 2, \dots, 4 \\ Y &= f_Y(\bar{L}(4), \bar{A}(4), U_Y) \end{aligned}$$

Is the true $P_{U,X}$ compatible with the SCM presented? Refer back to the second data structure of R Lab 1 for the true $P_{U,X}$.

2. Target causal parameter:

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

Is the target causal parameter (a parameter of $P_{U,X}$) identified (as a parameter of P_0) under the standard, point treatment randomization assumption/back door criteria? Why or why not?

3. **If the target parameter is not identified in the previous question, what are the alternative assumptions under which the parameter would be identified?**
4. **What is the corresponding statistical estimand, $\Psi(P_0)$, under these assumptions?**

5. **Bonus!** Suppose instead your target causal parameter is $E_{U,X}[Y_{a(1)}]$. What is the interpretation of this target parameter? What assumptions are needed for identifiability here? What is the statistical estimand?

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

1. SCM:

$U = (U_{L(t)}, U_{C(t)}, U_{A(t)}, U_{Y(t+1)}), t = 1, 2 \sim P_U$. Assume U s are jointly independent.
Structural equations, F :

$$\begin{aligned} L(1) &= f_{L(1)}(U_{L(1)}) \\ C(1) &= f_{C(1)}(L(1), U_{C(1)}) \\ A(1) &= f_{A(1)}(L(1), C(1), U_{A(1)}) \\ Y(2) &= f_{Y(2)}(L(1), C(1), A(1), U_{Y(2)}) \\ L(2) &= f_{L(2)}(L(1), C(1), A(1), Y(2), U_{L(2)}) \\ C(2) &= f_{C(2)}(\bar{L}(2), C(1), A(1), Y(2), U_{C(2)}) \\ A(2) &= f_{A(2)}(\bar{L}(2), \bar{C}(2), A(1), Y(2), U_{A(2)}) \\ Y(3) &= f_{Y(3)}(\bar{L}(2), \bar{C}(2), \bar{A}(2), Y(2), U_{Y(3)}) \end{aligned}$$

Is the true $P_{U,X}$ an element of the SCM presented? Refer back to the fourth data structure in R Lab 1 for the true $P_{U,X}$.

2. Target causal parameter:

$$\Psi^F(P_{U,X}) = E_{U,X}[Y(3)_{\bar{a}(2)=1, \bar{c}(2)=0}]$$

Is the target causal parameter (a parameter of $P_{U,X}$) identified (as a parameter of P_0) under the standard, point treatment randomization assumption/back door criteria? Why or why not?

3. **If the target parameter is not identified in the previous question, what are the alternative assumptions under which the parameter would be identified?**
4. **What is the corresponding statistical estimand, $\Psi(P_0)$, under these assumptions?**

3 For Your Project: Identification

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

1. Under what assumptions is the target causal parameter you came up with in the previous lab identified as a function of the observed data distribution?
2. What is your $\Psi(P_0)$, the statistical estimand?
3. Optional: confirm that in your simulation, the value of your estimand equals the value of your target causal parameter.

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?