R Lab 3 Part 1 - Understanding Time Dependent Confounding and Identifiability in Longitudinal Context

Advanced Topics in Causal Inference

Assigned: September 21, 2021
Lab due: September 28, 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:
Translate causal questions into target causal parameters for data structures 2 and 4.

---

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).

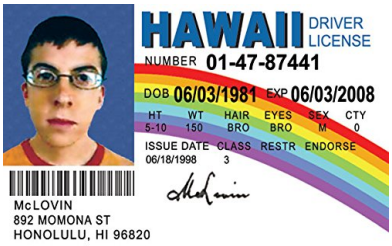

Fig. 1: Identification.

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 Us 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):

$$\begin{aligned} &\text{Exogenous variables:} \\ U_{L(t)} &\text{ for } t = 1, 2 \sim Uniform(min = 0, max = 1) \\ U_{A(t)} &\text{ for } t = 1, 2 \sim Uniform(min = 0, max = 1) \\ U_Y &\sim Uniform(min = 0, max = 1) \\ &\text{Structural equations } F \text{ and endogenous variables:} \\ L(1) &= \mathbb{I}[U_{L(1)} < 0.5] \\ A(1) &= \mathbb{I}[U_{A(1)} < expit(0.3 - L(1))] \\ L(2) &= \mathbb{I}[U_{L(2)} < expit(-2 + 1.8^* A(1) + 2^* L(1))] \\ A(2) &= \mathbb{I}[U_{A(2)} < expit(L(2) + L(1))] \\ Y &= \mathbb{I}[U_Y < expit(-3 + 1.3^* A(1) + 1.7^* A(2) + 1.3^* L(1) + 1.7^* L(2))] \end{aligned}$$

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

$$\Psi^F(P_{U,X}) = E_{U,X}[Y_{\bar{a}=1}] = 0.7921$$

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?**

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

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

$$\begin{aligned}
&\text{Permutation 1} = (L(1) = 1, L(2) = 1)\\
&\text{Permutation 2} = (L(1) = 1, L(2) = 0)\\
&\text{Permutation 3} = (L(1) = 0, L(2) = 1)\\
&\text{Permutation 4} = (L(1) = 0, L(2) = 0)
\end{aligned}$$

   For each of these permutations we need to compute:

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

   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:

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

   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 $P_0\big(\bar{L}(2) = 1\big)$ 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!*

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

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?**

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

(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:

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

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

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

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

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

Substitute in the constant values we're conditioning on:

$$= P\Big(U_Y < expit\big(-3 + 1.3^*1 + 1.7^*1 + 1.3^*1 + 1.7^*1\big)\Big)$$

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

$$= expit\big(-3 + 1.3^*1 + 1.7^*1 + 1.3^*1 + 1.7^*1\big)$$

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

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

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

$$P(L(0) = 1) = P\big(\mathbb{I}[U_{L(1)} < 0.5] = 1\big) = 0.5$$

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, $\Psi^{(3)}(P_0)$.

Solution:

1. Solve $\Psi^{(1)}(P_0)$ (recall that in this scenario, we are conditioning on $L(1)$ and $L(2)$):

```
> # set the seed
> set.seed(252)

> # number of times to generate
> n = 1000000

> # print data generating process for data structure 0
> print(generate_data0)

function(n){

  L1 = rbinom(n,1,0.5)
  A1 = rbinom(n,1,plogis(0.3-L1))
  L2 = rbinom(n,1,plogis(-2+1.8*A1+2*L1))
  A2 = rbinom(n,1,plogis(L2+L1))
  Y = rbinom(n,1,plogis(-3+1.3*A1+1.7*A2+1.3*L1+1.7*L2))

  O = data.frame(L1, A1, L2, A2, Y)
  return(O)

}
<bytecode: 0x7f80247ddc40>

> # save to dataframe ObsData0
> ObsData0 = generate_data0(n)

> # extract variables
> L1 = ObsData0$L1
> A1 = ObsData0$A1
> L2 = ObsData0$L2
> A2 = ObsData0$A2
> Y = ObsData0$Y

> # calculate Psi1(P0)
> Psi1.P0 = mean(Y[A1 == 1 & A2 ==1 & L1 == 1 & L2 == 1])*mean(L1 == 1 & L2 == 1) +
+   mean(Y[A1 == 1 & A2 ==1 & L1 == 1 & L2 == 0])*mean(L1 == 1 & L2 == 0) +
+   mean(Y[A1 == 1 & A2 ==1 & L1 == 0 & L2 == 1])*mean(L1 == 0 & L2 == 1) +
+   mean(Y[A1 == 1 & A2 ==1 & L1 == 0 & L2 == 0])*mean(L1 == 0 & L2 == 0)
> Psi1.P0

[1] 0.7477152
```

$\Psi^{(1)}(P_0) = 0.7477$. Here, $\Psi^{(1)}(P_0) \neq \Psi^F(P_{U,X})$; we have underestimated the target causal parameter.

2. Solve $\Psi^{(2)}(P_0)$ (recall that in this scenario, we are conditioning on $L(1)$ only):

```
> # calculate Psi2(P0)
> Psi2.P0 = mean(Y[A1 == 1 & A2 ==1 & L1 == 1])*mean(L1 == 1) +
+   mean(Y[A1 == 1 & A2 ==1 & L1 == 0])*mean(L1 == 0)
> Psi2.P0

[1] 0.8103919
```

$\Psi^{(2)}(P_0) = 0.8104$. Again, $\Psi^{(2)}(P_0) \neq \Psi^F(P_{U,X})$, and we have overestimated the target causal parameter.

3. Solve for $\Psi^{(3)}(P_0)$ (using longitudinal g-computation formula):

(a) Computationally:

```
> EY.11 = mean(Y[A1 == 1 & A2 ==1 & L1 == 1 & L2 == 1])*
+   mean(L2[A1 == 1 & L1 == 1])*
+   mean(L1)
> EY.10 = mean(Y[A1 == 1 & A2 ==1 & L1 == 1 & L2 == 0])*
+   (1 - mean(L2[A1 == 1 & L1 == 1]))*
+   mean(L1)
> EY.01 = mean(Y[A1 == 1 & A2 ==1 & L1 == 0 & L2 == 1])*
+   mean(L2[A1 == 1 & L1 == 0])*
+   (1 - mean(L1))
> EY.00 = mean(Y[A1 == 1 & A2 ==1 & L1 == 0 & L2 == 0])*
+   (1 - mean(L2[A1 == 1 & L1 == 0]))*
+   (1 - mean(L1))
> Psi3.P0 = EY.11 + EY.10 + EY.01 + EY.00
> Psi3.P0

[1] 0.792322
```

(b) Analytically:
Continuing for all other permutations of $(L(1), L(2))$...

- Case where $(L(1) = 1, L(2) = 1)$):

$$\begin{aligned} &E[Y|\bar{A} = 1, L(1) = 1, L(2) = 1]P(L(2) = 1|A(1) = 1, L(1) = 1)P(L(1) = 1) \\ &= expit(3)^* expit(1.8)^* 0.5 \\ &= 0.4087 \end{aligned}$$

- Case where $(L(1) = 1, L(2) = 0)$):

$$\begin{aligned} &E[Y|\bar{A} = 1, L(1) = 1, L(2) = 0]P(L(2) = 0|A(1) = 1, L(1) = 1)P(L(1) = 1) \\ &= expit(1.3)^*(1 - expit(1.8))^* 0.5 \\ &= 0.0557 \end{aligned}$$

- Case where $(L(1) = 0, L(2) = 1)$):

$$\begin{aligned} &E[Y|\bar{A} = 1, L(1) = 0, L(2) = 1]P(L(2) = 1|A(1) = 1, L(1) = 0)P(L(1) = 0) \\ &= expit(1.7)^* expit(-0.2)^* 0.5 \\ &= 0.1903 \end{aligned}$$

- Case where $(L(1) = 0, L(2) = 0)$):

$$\begin{aligned} &E[Y|\bar{A} = 1, L(1) = 0, L(2) = 0]P(L(2) = 0|A(1) = 1, L(1) = 0)P(L(1) = 0) \\ &= expit(0)^*(1 - expit(-0.2))^* 0.5 \\ &= 0.1375 \end{aligned}$$

Adding these components together we get $\Psi^{(3)}(P_0) = 0.7922$

We see that $\Psi^{(3)}(P_0)$, which uses the longitudinal g-computation formula, does equal the value of our target causal estimand (assume any difference is due to rounding error). Under the sequential randomization and positivity assumption, this statistical estimand identifies our causal parameter of interest.

Note that the above does not constitute a proof of this identification result - identification implies that this equality will hold for every $P_{U,X}$ compatible with the SCM, and every P_0 implied by this $P_{U,X}$. We have just shown that it holds for one such $P_{U,X}$.

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 Us 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?**

Solution:

1. $P_{U,X}$ is described by the SCM. An SCM is a model on the set of possible data generating processes. The data generating process (i.e., $P_{U,X}$) we wrote down in `R` Lab 1 is an element (a specific instance) of our SCM written above.

2. The point treatment backdoor criteria here would require one set of covariates that blocks backdoor paths from the joint intervention (A, Δ) to Y *and* does not include descendants of (A, Δ). Our two options are:

   - Conditioning on W. Here, the backdoor criteria fails because there is a backdoor path from $\Delta \leftarrow L \rightarrow Y$ and $(Y_{a=0,\Delta=1} \not\perp\!\!\!\perp (A, \Delta)|W)$. Thus, our target causal parameter $E_{U,X}[Y_{a=0,\Delta=1}] \neq E_0[E_0[Y|A = 0, \Delta = 1, W]]$. Informally, if we fail to adjust for L, we have informative missingness.
   - Conditioning on W and L. Here, the backdoor criteria fails because L is affected by A. Thus, our target causal parameter $E_{U,X}[Y_{a=0,\Delta=1}] \neq E_0[E_0[Y|A = 0, \Delta = 1, W, L]]$. This is because the right hand side of the inequality is integrating with respect to the observed distribution of L, rather than the counterfactual distribution under an intervention to set $A = 0$. Informally, if we just condition on L in a single regression, we lose part of the effect of A on Y.

   Thus, the target causal parameter is *not* identified using the point treatment g-computation formula.

3. Alternative assumptions sufficient to identify the target causal parameter:

   - Sequential randomization assumption (and corresponding sequential backdoor criteria):

$$\begin{aligned} Y_{a=0,\Delta=1} &\perp A|W \\ Y_{a=0,\Delta=1} &\perp \Delta|W, L, A = 0 \end{aligned}$$

- Positivity assumption

$$P_0(A = 0|W = w) > 0 - a.e.$$
$$P_0(\Delta = 1|W = w, L = l, A = 0) > 0 - a.e.$$

4. Under the above assumptions, the target causal parameter is equal to a function of the observed data distribution. Specifically:

$$E_{U,X}[Y_{a=0,\Delta=1}] = \sum_{w,l} E_0[Y|A = 0, \Delta = 1, W = w, L = l]P_0(L = l|A = 0, W = w)P_0(W = w)$$

3 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. Was there enough group time to work on each question?
3. What did you learn in this lab?
4. Do you think that this lab met the goals listed at the beginning?
5. What else would you have liked to review? What would have helped your understanding?
6. Any other feedback?