R Assignment 1:
Causal Parameters & Simulations in R

Laura B. Balzer

Biostat683 - Intro. to Causal Inference

Assigned: September 20, 2021
Write-ups due: Uploaded to your personal GoogleDrive folder by October 4, 2021 by 2:30pm. 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. Use of RMarkdown is strongly encouraged; please see the template on GoogleDrive.

1 Background Story

Suppose we are interested in the causal effect of ready-to-use therapeutic food (RUTF) on recovery from undernutrition in a resource-limited country. RUTF is peanut butter-type paste, fortified with milk proteins and essential nutrients, and does not require water for use (WHO, 2007). We propose a study to contrast the effect of RUTF with the standard supplement on weight gain over two months among school-aged children.

Suppose we only have two pre-exposure covariates. Specifically, $W1$ is an indicator equaling 1 if the child has access to potable water. Likewise, $W2$ is an indicator equaling 1 if the child suffered from an infectious disease within the two weeks prior to the study initiation. The intervention A is also an indicator equaling 1 if the child received RUTF and 0 if the child received the standard supplement. Finally, the outcome Y represents the child's weight gain in pounds at the study close.

The above study can be translated into the following structural causal model (SCM) $\mathcal{M}^*$:

- Endogenous nodes: $X = (W1, W2, A, Y)$
- Background variables: $U = (U_{W1}, U_{W2}, U_A, U_Y) \sim \mathbb{P}_U$
- Structural equations $\mathcal{F}$:

$$\begin{aligned} W1 &= f_{W1}(U_{W1}) \\ W2 &= f_{W2}(W1, U_{W2}) \\ A &= f_A(W1, W2, U_A) \\ Y &= f_Y(W1, W2, A, U_Y) \end{aligned}$$

2 Steps 1-2 of the Roadmap

1. **Step 1: Causal model representing real knowledge**
   (a) Draw the accompanying directed acyclic graph (DAG).
   (b) Are there any exclusion restrictions? Recall we are working with recursive (time-ordered) structural causal models.

(c) Are there any independence assumptions on the distribution of unmeasured factors $\mathbb{P}_U$?

2. **Step 2: Counterfactuals & causal parameter**

(a) Define the counterfactual outcomes of interest with formal notation and in words.

(b) How are counterfactuals derived?

(c) Suppose we are interested in the average treatment effect. Specify the target causal parameter. Use formal notation as well as explain in words.

3 A specific data generating process

Now, consider a particular data generating process, one of many compatible with $\mathcal{M}^*$. Suppose that the each of the background factors is drawn independently from following distributions:

$$\begin{aligned} U_{W1} &\sim Uniform(0,1) \\ U_{W2} &\sim Uniform(0,1) \\ U_A &\sim Uniform(0,1) \\ U_Y &\sim Normal(\mu = 0, \sigma^2 = 0.3^2) \end{aligned}$$

Given the background U, the endogenous variables are deterministically generated as

$$\begin{aligned} W1 &= \mathbb{I}[U_{W1} < 0.2] \\ W2 &= \mathbb{I}[U_{W2} < logit^{-1}(0.5 \times W1)] \\ A &= \mathbb{I}[U_A < logit^{-1}(W1 \times W2)] \\ Y &= 4 \times A + 0.7 \times W1 - 2 \times A \times W2 + U_Y \end{aligned}$$

Recall the $logit^{-1}$ is the inverse-logit:

$$logit^{-1}(x) = \frac{exp(x)}{1 + exp(x)}$$

and given by the `plogis` function in `R`.

3.1 Closed form evaluation of the target causal parameter

Evaluate the target causal parameter $\Psi^*(\mathbb{P}^*)$ in closed form (i.e., by hand) for this data generating process.
Hints: In this particular data generating system (one of many compatible with the SCM), the expectation of the counterfactual outcome is a linear function of the treatment level a, the pre-exposure covariates $(W1, W2)$ and random error U_Y:

$$\mathbb{E}^*(Y_a) = \mathbb{E}^*[f_Y(W1, W2, a, U_Y)] = \mathbb{E}^*\big[4 \times a + 0.7 \times W1 - 2 \times a \times W2 + U_Y\big]$$

The marginal distribution of $W1$ (access to potable water) is Bernoulli with probability 0.20:

$$\mathbb{P}^*(W1 = 1) = \mathbb{E}^*(W1) = 0.20$$

The conditional expectation of $W2$ (presence or absence of an infectious disease), given $W1$, is given by

$$\mathbb{P}^*(W2 = 1 \mid W1) = \mathbb{E}^*(W2 \mid W1) = logit^{-1}(0.5 \times W1)$$

3.2 Translating this data generating process into simulations, generating counterfactual outcomes, and evaluating the target causal parameter.

1. **First set the seed to 252.**
2. **Set `n=50,000` as the number of i.i.d. draws from the data generating process.**
3. **Simulate the background factors U.** Note the syntax for `rnorm`.
4. **Evaluate the structural equations $\mathcal{F}$ to deterministically generate the endogenous nodes X.** Recall the $logit^{-1}$ function is given by the `plogis` function in `R`.
5. **Intervene to set the supplement to RUTF ($A = 1$) and generate counterfactual outcomes Y_1 for n units. Then intervene to set the supplement to the standard ($A = 0$) and generate counterfactual outcomes Y_0 for n units.**
6. **Create a data frame `X` to hold the values of the endogenous factors $(W1, W2, A, Y)$ and the counterfactual outcomes Y_1 and Y_0. The rows are the n children and the columns are their characteristics. Use the `head` and `summary` to examine the resulting data. Does the counterfactual value Y_a equal the observed Y when $A = a$?**
7. **Using these simulations, evaluate the causal parameter $\Psi^*(\mathbb{P}^*)$ for this population of 50,000 units.**
8. **Interpret $\Psi^*(\mathbb{P}^*)$.**

4 Defining the target causal parameter with a working MSM

Now suppose we are interested in knowing if age (in years) V modifies the effect of RUTF A on weight gain Y. As before, $W1$ is an indicator of access to potable water and $W2$ is an indicator of having an infectious disease within two weeks of the study initiation.

Consider the following SCM $\mathcal{M}^*$:

- Endogenous nodes: $X = (V, W1, W2, A, Y)$
- Background nodes: $U = (U_V, U_{W1}, U_{W2}, U_A, U_Y) \sim \mathbb{P}_U$
- Structural equations $\mathcal{F}$:

$$\begin{aligned} V &= f_V(U_V) \\ W1 &= f_{W1}(U_{W1}) \\ W2 &= f_{W2}(V, W1, U_{W2}) \\ A &= f_A(V, W1, W2, U_A) \\ Y &= f_Y(V, W1, W2, A, U_Y) \end{aligned}$$

- We have made an exclusion restriction that age V does not effect access to potable water $W1$.

Let us summarize how the counterfactual outcome changes as a function of the intervention and age with the following *working* marginal structural model (MSM):

$$\begin{aligned} \beta^* &= argmin_\beta \, \mathbb{E}^* \left[\sum_{a \in \mathcal{A}} \left(Y_a - m(a, V|\beta) \right)^2 \right] \\ m(a, V|\beta) &= \beta_0 + \beta_1 a + \beta_2 V + \beta_3 a \times V \end{aligned}$$

Then the target parameter is defined as a projection of the true causal curve $\mathbb{E}^*(Y_a|V)$ onto a working model $m(a, V|\beta)$. In other words, the causal parameters are the values of the β coefficients that minimize the sum of

squared residuals between the counterfactuals Y_a and the model predictions $m(a, V|\beta)$ for all possible exposure levels $a \in \mathcal{A}$.

Based on our knowledge of the data generating system, as represented in $\mathcal{M}^*$, this working MSM with an interaction term may or may not be a good summary of how the effect of RUTF on the counterfactual average weight gain is modified by age.

4.1 A specific data generating process:

Consider a new data generating process (one of many compatible with the SCM). Suppose that the each of the background factors is drawn independently from following distributions:

$$\begin{aligned} U_V &\sim Uniform(0,3) \\ U_{W1} &\sim Uniform(0,1) \\ U_{W2} &\sim Uniform(0,1) \\ U_A &\sim Uniform(0,1) \\ U_Y &\sim Normal(\mu = 0, \sigma^2 = 0.1^2) \end{aligned}$$

Given the background factors U, the endogenous variables X are deterministically generated as

$$\begin{aligned} V &= 2 + U_V \\ W1 &= \mathbb{I}[U_{W1} < 0.2] \\ W2 &= \mathbb{I}[U_{W2} < logit^{-1}(0.5^* W1)] \\ A &= \mathbb{I}[U_A < logit^{-1}(W1^* W2 + V/5)] \\ Y &= 4^* A + 0.7^* W1 - 2^* A^* W2 + .3^* V - .3^* A^* V + U_Y \end{aligned}$$

1. **For $n = 5,000$ children, generate the background factors U and the pre-exposure covariates $(V, W1, W2)$. Then set $A = 1$ to generate the counterfactual weight gain under RUTF Y_1. Likewise, set $A = 0$ to generate the counterfactual weight gain under the standard supplement Y_0.**

2. **Create a data frame `X.msm` consisting of age V, the set treatment levels a, and the corresponding outcomes Y_a.**

$$X_{MSM} = (V, a, Y_a) = \begin{pmatrix} V(1) & 1 & Y_1(1) \\ V(2) & 1 & Y_1(2) \\ \vdots & \vdots & \vdots \\ V(n) & 1 & Y_1(n) \\ V(1) & 0 & Y_0(1) \\ V(2) & 0 & Y_0(2) \\ \vdots & \vdots & \vdots \\ V(n) & 0 & Y_0(n) \end{pmatrix}$$

   where $V(i)$ and $Y_a(i)$ denote the age and counterfactual outcome for the i^{th} participant. See `R` lab 1 for a similar example.

3. **Evaluate the target causal parameter.** We have defined the target parameter using the least square projection (i.e., with the L2 loss function). Use the `glm` function to fit the coefficients of the working MSM. Specifically, regress the counterfactual outcomes Y_a on a and V according to the working MSM. Be sure to specify the argument: `data=X.msm`.

4. **Interpret the results.**

5. **Bonus:** Plot of the counterfactual outcomes Y_a as a function of age (V) and treatment group (a).

References

World Health Organization (WHO), World Food Programme (WFP), United Nations System Standing Committee on Nutrition (SCN), and United Nations Children's Fund (UNICEF). *Community-based management of severe acute malnutrition.* WHO/WFP/SCN/UNICEF, Geneva/Rome/Geneva/New York, 2007.