R Assignment 1 - Causal Parameters & Simulations in R

Introduction to Causal Inference

Write-up: 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. You need to bring your own completed assignment to class.

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-intervention 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 termination.

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

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

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

1. Draw the accompanying DAG.
2. Are there any exclusion restrictions?
3. Are there any independence assumptions?
4. Define the counterfactual outcomes of interest with formal notation and in words. How are counterfactuals derived?
5. Suppose we are interested in the average treatment effect. Specify the target causal parameter. Use formal notation as well as explain in words.
6. Suppose the observed data consist of n independent, identically distributed (i.i.d) draws of the random variable $O = (W1, W2, A, Y)$. Specify the link between the SCM and the observed data? What restrictions, if any, does the SCM place on the allowed distributions for the observed data? What notation do we use to denote the true (but unknown) distribution of the observed data and the statistical model?
7. Using the backdoor criteria, assess identifiability. If the target causal parameter is not identified, under what assumptions would it be? What notation is used to denote the original SCM augmented with additional assumptions needed for identifiability?

8. Specify the target parameter of the observed data distribution (the statistical estimand).
9. What is the relevant positivity assumption? Is it reasonable here?

2 Highly Recommended Bonus: Identifying the Mean Outcome Under a Dynamic Intervention

Note: This section is a bonus, but may be helpful for getting your head around about why the backdoor criterion allows us to identify our causal parameter with a statistical parameter. This problem considers dynamic treatment rules, but the same general arguments also give identifiability for static treatment rules. Feel free to come to office hours if you want to work through the problem with one of your GSIs.

Suppose the investigators are also interested in the population mean outcome if, possibly contrary to fact, the dynamic treatment d is given in the population, where

$$\begin{aligned} d(W2) &= I(W2 = 1) \\ &= \begin{cases} 1, & \text{if the child suffered from an infectious disease two weeks prior to study initiation} \\ 0, & \text{otherwise.} \end{cases} \end{aligned}$$

That is, the investigators are interested in learning about the causal parameter $\Psi_d^F(P_{U,X}) = E_{U,X}[Y_d]$. If we assume that W satisfies the backdoor criterion for the effect of A on Y, then this will imply a randomization assumption for the rule d:

$$Y_d \perp\!\!\!\perp A \mid W_1, W_2$$

i.e. Y_d is independent of A given W_1 and W_2. We will also assume the same positivity assumption you gave for Question 9 in Section 1. The objective of this exercise is to understand why $\Psi_d^F(P_{U,X}) = \Psi_d(P_0)$ under the randomization assumption, where

$$\Psi_d(P_0) = \sum_{w1,w2} E_0\left[Y|A = d(w2), W1 = w1, W2 = w2\right] P_0\left(W1 = w1, W2 = w2\right).$$

We will give you the derivation of this equality below, and then we will ask you to justify each of the equalities in the derivation using both properties of random variables and a translation of those properties to the RUTF data structure. We have that

$$\begin{aligned} \Psi_d^F(P_{U,X}) &= E_{U,X}[Y_d] \\ &= \sum_{w1,w2} E_{U,X}[Y_d|W1 = w1, W2 = w2] P_{U,X}(W1 = w1, W2 = w2) && (1) \\ &= \sum_{w1,w2} \sum_y y P_{U,X}\left(Y_d = y|W1 = w1, W2 = w2\right) P_{U,X}(W1 = w1, W2 = w2) && (\star) \\ &= \sum_{w1,w2} \sum_y y P_{U,X}\left(Y_d = y|A = d(w2), W1 = w1, W2 = w2\right) P_{U,X}(W1 = w1, W2 = w2) && (2) \\ &= \sum_{w1,w2} \sum_y y P_0\left(Y = y|A = d(w2), W1 = w1, W2 = w2\right) P_{U,X}(W1 = w1, W2 = w2) && (3) \\ &= \sum_{w1,w2} E_0[Y|A = d(w2), W1 = w1, W2 = w2] P_{U,X}(W1 = w1, W2 = w2) && (\star) \\ &= \sum_{w1,w2} E_0[Y|A = d(w2), W1 = w1, W2 = w2] P_0(W1 = w1, W2 = w2) && (4) \\ &= \Psi_d(P_0), \end{aligned}$$

where each $(\star)$ holds by the definition of conditional expectation. Below we ask you to justify labeled equalities (1) through (4). Note that we have implicitly used the positivity assumption in (2) and all of the subsequent

equalities, since positivity ensures the conditional expectations and probabilities make sense – it is impossible to calculate the average outcome in a strata which does not contain any individuals (occurs with probability 0) in the target population!

1. Explain why (1) holds using properties of conditional expectations. Given access to the full population and the ability to implement intervention d, what does (1) tell you about how you could compute $E_{U,X}[Y_d]$?
   `Hint: Look up the law of total expectation.`
2. Explain why (2) holds using properties of conditional expectations and the fact that $Y_d \perp\!\!\!\perp A|W_1, W_2$ under our convenience assumptions for the backdoor criterion made in Question 7 of Section 1.
   `Note: No need to explain` $Y_d \perp\!\!\!\perp A|W_1, W_2$ `in the context of the study since you have already discussed the assumptions needed for the backdoor criterion to hold, and the backdoor criterion implies` $Y_d \perp\!\!\!\perp A|W_1, W_2$.
3. Explain why (3) holds. What does this mean in terms of the RUTF example?
   `Hint: Recall that` $Y_d = f_Y(W1, W2, d(W2), U_Y)$ `and` $Y = f_Y(W1, W2, A, U_Y)$.
4. Explay why (4) holds. What does this mean in terms of the RUTF example?
   `Hint: Does our intervention affect our baseline covariates?`

3 A specific data generating process

Now, consider a particular data generating process $P_{U,X}$, one of many compatible with $\mathcal{M}^F$. Suppose that the each of the exogenous 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 exogenous U, the endogenous variables are deterministically generated as

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

Recall the *expit* function is the inverse of the logit function: $expit(x) = 1/(1 + e^{-x})$.

1. Evaluate the target causal parameter $\Psi^F(P_{U,X})$ in closed form 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-intervention covariates $(W1, W2)$ and random error U_Y:

   $$E_{U,X}(Y_a) = E_{U,X}\left[4^*a + 0.7^*W1 - 2^*a^*W2 + U_Y\right]$$

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

   $$P_{U,X}(W1 = 1) = E_{U,X}(W1) = 0.20$$

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

   $$P_{U,X}(W2 = 1||W1) = E_{U,X}(W2|W1) = expit(0.5^*W1)$$

   By the tower rule, the marginal expectation of $W2$ is given by

   $$\begin{aligned} E_{U,X}(W2) &= \sum_{w1} E_{U,X}(W2|W1 = w1) P_{U,X}(W1 = w1) \\ &= E_{U,X}(W2|W1 = 1) P_{U,X}(W1 = 1) + E_{U,X}(W2|W1 = 0) P_{U,X}(W1 = 0) \end{aligned}$$

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

3.1 Translating this data generating process for $P_{U,X}$ into simulations, generating counterfactual outcomes and evaluating the target causal parameter.

1. First set the seed to 252.
2. Set `n=5000` 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 F to deterministically generate the endogenous nodes X. Recall the *expit* 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.
7. Evaluate the causal parameter $\Psi^F(P_{U,X})$.

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 $\mathcal{M}^F$:

- Endogenous nodes: $X = (V, W1, W2, A, Y)$
- Exogenous nodes: $U = (U_V, U_{W1}, U_{W2}, U_A, U_Y) \sim P_U$
- Structural equations 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):

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

Then the target parameter is defined as a projection of the true causal curve $E_{U,X}(Y_a|V)$ onto a working model $m(a, V|\beta)$. In other words, the causal parameter is the value of the β coefficients that minimize the sum of squared residuals between the counterfactuals Y_a and the model $m(a, V|\beta)$ for all possible exposure levels $a \in \mathcal{A}$.

Based on our knowledge of the data generating system, as represented in $\mathcal{M}^F$, a linear 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 exogenous 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 exogenous U, the endogenous variables are deterministically generated as

$$\begin{aligned} V &= 2 + U_V \\ W1 &= \mathbb{I}[U_{W1} < 0.2] \\ W2 &= \mathbb{I}[U_{W2} < expit(0.5^* W1)] \\ A &= \mathbb{I}[U_A < expit(W1^* W2 + V/5)] \\ Y &= 2^* A + 0.3^* W1 + 2^* A^* W2 + 0.5^* A^* V + U_Y \end{aligned}$$

1. For $n = 5000$ children, generate the exogenous factors U and the pre-intervention 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} subject. 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.

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.