# 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  $\mathcal{F}$ :

$$W1 = f_{W1}(U_{W1})$$

$$W2 = f_{W2}(W1, U_{W2})$$

$$A = f_A(W1, W2, U_A)$$

$$Y = f_Y(W1, W2, A, U_Y)$$

# 2 Steps 1-2 of the Roadmap

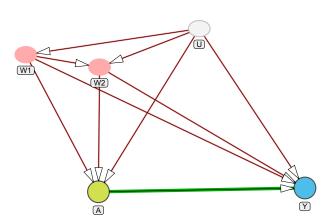
- 1. Step 1: Causal model representing real knowledge
  - (a) Draw the accompanying directed acyclic graph (DAG).
  - (b) Are there any exclusion restrictions? Recall we are working with recursive (time-ordered) structural causal models.

(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.

#### Solution:



Solution Fig. 1: Directed Acyclic Graph for the study of RUTF on weight gain. Recall W1 represents access to potable water, W2 represents recent history of an infectious disease, A is the exposure (nutritional supplementation) and Y is the outcome (weight gain in pounds).

### 1. Step1

- (a) See Solution Fig. 1.
- (b) There are no exclusion restrictions. All endogenous nodes are affected by the nodes preceding them temporally. We are assuming access to potable water W1 may affect presence of infectious disease within two weeks of study initiation W2.
- (c) There are no independence assumptions explicitly specified in the background above. If you made any independence assumptions on  $\mathbb{P}_U$ , you should justify them.

#### 2. Step 2

- (a) The counterfactual outcomes of interest are  $Y_a: a \in \mathcal{A} = \{0,1\}$ .  $Y_1$  is a child's weight gain if, possibly contrary to fact, they received RUTF A = 1.  $Y_0$  is a child's weight gain if, possibly contrary to fact, they received the standard supplement A = 0.
- (b) Counterfactuals are derived by intervening on the SCM to set A = a. The distribution of counterfactuals, denoted  $\mathbb{P}^*$ , is implied by the distribution of background and endogenous factors, denoted  $\mathbb{P}_U$ , and set of structural equations, denoted  $\mathcal{F}$ . The SCM  $\mathcal{M}^*$  gives us a model for the set of possible counterfactual distributions.
- (c) The target causal parameter is the difference in the expected counterfactual weight gain if all children were given RUTF and the expected counterfactual weight gain if all children were given the standard supplement:

$$\Psi^*(\mathbb{P}^*) = \mathbb{E}^*(Y_1) - \mathbb{E}^*(Y_0)$$
  
=  $\mathbb{E}^*[f_Y(W_1, W_2, 1, U_Y)] - \mathbb{E}^*[f_Y(W_1, W_2, 0, U_Y)]$ 

In the second line, we have replaced the counterfactual outcome  $Y_a$  with the corresponding (post-intervention) structural equation.

# 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:

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

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

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

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  $U_Y$ :

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

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

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

Solution: For this data generating process, the counterfactual mean outcome is

$$\begin{split} \mathbb{E}^*[Y_a] &= \mathbb{E}^*[4a + 0.7W1 - 2a \times W2 + U_Y] \\ &= 4a + 0.7\mathbb{E}^*[W1] - 2a \times \mathbb{E}^*[W2] + \mathbb{E}^*[U_Y] \end{split}$$

The marginal expectation of W1 is  $\mathbb{E}^*[W1] = 0.20$ . The marginal expectation of W2 is

$$\begin{split} \mathbb{E}^*[W2] &= \mathbb{E}^*(W2|W1=1)\mathbb{P}^*(W1=1) + \mathbb{E}^*(W2|W1=0)\mathbb{P}^*(W1=0) \\ &= logit^{-1}(0.5\times1)*0.20 + logit^{-1}(0.5\times0)*(1-0.20) \end{split}$$

The expectation of the background factor  $U_Y$  is zero.

Plugging these values into the target causal parameter, we have

$$\begin{split} \Psi^*(\mathbb{P}^*) = & \mathbb{E}^*(Y_1) - \mathbb{E}^*(Y_0) \\ = & 4 \times 1 + 0.7 \times \mathbb{E}^*[W1] - 2 \times 1 \times \mathbb{E}^*[W2] + \mathbb{E}^*[U_Y] \\ & - \left(4 \times 0 + 0.7 \times \mathbb{E}^*[W1] - 2 \times 0 \times \mathbb{E}^*[W2] + \mathbb{E}^*[U_Y]\right) \\ = & 4 \times 1 + 0.7 \times 0.20 - 2 \times 1 \times \left[logit^{-1}(0.5 \times 1) * 0.20 + logit^{-1}(0.5 \times 0) * (1 - 0.20)\right] - \left(0.7 \times 0.20\right) \\ \approx & 3.09 - 0.14 \end{split}$$

Therefore, the average treatment effect is

$$\Psi^*(\mathbb{P}^*) = \mathbb{E}^*(Y_1) - \mathbb{E}^*(Y_0) \approx 3.09 - 0.14 = 2.95$$

> # corresponding R code for E^\*[Y\_1] > 4 + .7\*.2 -2\*(plogis(0.5)\*.2 + plogis(0)\*.8)

[1] 3.091016

Alternatively, we could have evaluated the target causal parameter as follows. The expected counterfactual outcome under the intervention to set A = a is

$$\begin{split} \mathbb{E}^*(Y_a) &= \mathbb{E}^* \left[ \mathbb{E}^*(Y_a | W1, W2) \right] \\ &= \sum_{w1, w2} \left[ \mathbb{E}^*(Y_a | W1 = w1, W2 = w2) \times \mathbb{P}^*(W1 = w1, W2 = w2) \right] \\ &= \sum_{w1, w2} \left[ \mathbb{E}^*(Y_a | W1 = w1, W2 = w2) \times \mathbb{P}^*(W2 = w2 | W1 = w1) \times \mathbb{P}^*(W1 = w1) \right] \end{split}$$

We know conditional mean of the counterfactual  $Y_a$  given W1 and W2 is

$$\mathbb{E}^{*}(Y_{a}|W1, W2) = 4 \times a + 0.7 \times W1 - 2 \times a \times W2$$

We also know the conditional distribution of W2, given W1. Specifically, we know

$$\mathbb{P}^*(W2 = 1|W1 = w1) = logit^{-1}(0.5w1)$$
$$\mathbb{P}^*(W2 = 0|W1 = w1) = 1 - logit^{-1}(0.5w1)$$

Therefore, the expected value of the counterfactual outcome if all children received RUTF (A = 1) is

$$\begin{split} \mathbb{E}^*(Y_1) &= \mathbb{E}^*(Y_1|W1=1,W2=1)\mathbb{P}^*(W2=1|W1=1)\mathbb{P}^*(W1=1) \\ &+ \mathbb{E}^*(Y_1|W1=0,W2=1)\mathbb{P}^*(W2=1|W1=0)\mathbb{P}^*(W1=0) \\ &+ \mathbb{E}^*(Y_1|W1=1,W2=0)\mathbb{P}^*(W2=0|W1=1)\mathbb{P}^*(W1=1) \\ &+ \mathbb{E}^*(Y_1|W1=0,W2=0)\mathbb{P}^*(W2=0|W1=0)\mathbb{P}^*(W1=0) \\ &= (4*1+0.7*1-2*1*1)logit^{-1}(0.5*1)0.2 \\ &+ (4*1+0.7*0-2*1*1)logit^{-1}(0.5*0)0.8 \\ &+ (4*1+0.7*1-2*1*0)(1-logit^{-1}(0.5*1))0.2 \\ &+ (4*1+0.7*0-2*1*0)(1-logit^{-1}(0.5*0))0.8 \\ &\approx 3.09 \end{split}$$

Likewise, the expected value of the counterfactual outcome if all children received that standard supplement (A = 0) is

$$\mathbb{E}^{*}(Y_{0}) = (4 * 0 + 0.7 * 1 - 2 * 0 * 1) logit^{-1}(0.5 * 1)0.2$$

$$+ (4 * 0 + 0.7 * 0 - 2 * 0 * 1) logit^{-1}(0.5 * 0)0.8$$

$$+ (4 * 0 + 0.7 * 1 - 2 * 0 * 0)(1 - logit^{-1}(0.5 * 1))0.2$$

$$+ (4 * 0 + 0.7 * 0 - 2 * 0 * 0)(1 - logit^{-1}(0.5 * 0))0.8$$

$$= 0.14$$

- 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  $\Psi^*(\mathbb{P}^*)$ .

```
Solution:
```

```
> # 1. set seed and number of draws
> set.seed(252)

> # 2. set the number of draws n=50,000
> n = 50000

> # 3. sample the background factors U
> U.W1<- runif(n, 0, 1)
> U.W2<- runif(n, 0, 1)
> U.A<- runif(n, 0, 1)
> U.Y<- rnorm(n, 0, 0.3)</pre>
```

```
> # 4. generate the endogenous factors X
> W1 <- as.numeric( U.W1 < 0.2)
> W2 <- as.numeric( U.W2 < plogis(0.5*W1) )
> A <- as.numeric( U.A < plogis(W1*W2))</pre>
> Y < -4*A + 0.7*W1 - 2*A*W2 + U.Y
> # 5. generate the counterfactuals for large n
> Y.1<- 4*1 + 0.7*W1 - 2*1*W2 + U.Y
> Y.O<- 4*0 + 0.7*W1 - 2*0*W2 + U.Y
> # 6. create data frame X
> X<- data.frame(W1, W2, A, Y, Y.1, Y.0)
> head(X)
 W1 W2 A
                     Y
                            Y.1
  0 0 0 -0.188412377 3.811588 -0.188412377
 0 0 0 0.007554149 4.007554 0.007554149
3 0 1 0 0.485863200 2.485863 0.485863200
  0 0 0 0.207939977 4.207940 0.207939977
5 0 1 0 -0.189080671 1.810919 -0.189080671
6 0 1 0 -0.017563026 1.982437 -0.017563026
> summary(X)
                        W2
                                                          Y
       :0.0000
                         :0.0000
                                          :0.0000
Min.
                                                           :-1.15726
                 Min.
                                   Min.
                                                    Min.
 1st Qu.:0.0000
                 1st Qu.:0.0000
                                   1st Qu.:0.0000
                                                    1st Qu.: 0.09119
Median :0.0000
                 Median :1.0000
                                   Median :1.0000
                                                    Median: 1.67785
                         :0.5254
                                          :0.5292
Mean :0.1976
                 Mean
                                   Mean
                                                    Mean : 1.67314
 3rd Qu.:0.0000
                 3rd Qu.:1.0000
                                   3rd Qu.:1.0000
                                                    3rd Qu.: 3.04173
       :1.0000
                        :1.0000
 Max.
                 Max.
                                   Max.
                                          :1.0000
                                                    Max.
                                                          : 5.66234
                       Y.0
     Y.1
        :0.8427
Min.
                 Min.
                         :-1.15726
                 1st Qu.:-0.15171
 1st Qu.:2.0867
Median :2.9530
                 Median: 0.08838
Mean :3.0871
                 Mean : 0.13794
 3rd Qu.:4.0423
                  3rd Qu.: 0.38180
Max.
       :5.6670
                 Max.
                       : 1.94849
> # 7. evaluate the causal parameter
> Psi.star<- mean(Y.1) - mean(Y.0)</pre>
> Psi.star
[1] 2.94912
```

The counterfactual mean weight gain would be 2.95 pounds higher if all children were given RUTF as opposed to the standard supplement.

# 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  $\mathcal{F}$ :

$$V = f_V(U_V)$$

$$W1 = f_{W1}(U_{W1})$$

$$W2 = f_{W2}(V, W1, U_{W2})$$

$$A = f_A(V, W1, W2, U_A)$$

$$Y = f_Y(V, W1, W2, A, U_Y)$$

- 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):

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

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  $\beta$  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:

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

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

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

- 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  $Y_a$ .

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

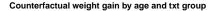
where V(i) and  $Y_a(i)$  denote the age and counterfactual outcome for the  $i^{th}$  participant. See R lab 1 for a similar example.

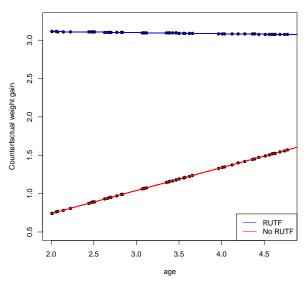
- 3. Evaluate the target causal parameter. We have defined the target parameter using the least square projection (i.e., with the L2 loss function). Use the glm function to fit the coefficients of the working MSM. Specifically, regress the counterfactual outcomes  $Y_a$  on a and V according to the working 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).

```
Solution:
> # set the number of draws n=5000
> n = 5000
> # sample the background factors U
> U.V<- runif(n, 0, 3)
> U.W1<- runif(n, 0, 1)
> U.W2<- runif(n, 0, 1)
> U.A<- runif(n, 0, 1)
> U.Y<- rnorm(n, 0, 0.1)
> # generate the pre-exposure covariates and the counterfactuals
> V <- 2 + U.V
> W1 <- as.numeric( U.W1 < 0.2)
> W2 <- as.numeric( U.W2 < plogis(0.5*W1) )</pre>
> Y.1 <- 4*1 + 0.7*W1 - 2*1*W2 + .3*V - .3*1*V + U.Y
> Y.O <- 4*0 + 0.7*W1 - 2*0*W2 + .3*V - .3*0*V + U.Y
> mean(Y.1)
[1] 3.091404
> mean(Y.0)
```

```
[1] 1.188447
> # 2. create data frame X.msm
> V < - rep(V,2)
> A<- c( rep(1,n), rep(0,n))
> Ya<- c(Y.1, Y.0)
> X.msm<- data.frame(V, A, Ya)
> head(X.msm)
        V A
                Ya
1 2.431051 1 4.619084
2 2.898271 1 4.520995
3 2.087969 1 2.201596
4 3.034100 1 4.044636
5 2.060597 1 2.552383
6 3.503814 1 1.957732
> tail(X.msm)
            V A
9995 3.631921 0 1.0303005
9996 2.984406 0 1.0241824
9997 3.224092 0 0.8730339
9998 2.966538 0 1.5122934
9999 3.611784 0 1.0968881
10000 2.405089 0 0.6879200
> # 3. evaluate the causal parameter
> #We can use ordinary least squares regression to obtain betas.
> workMSM<- glm(Ya ~A + V+ A*V, data=X.msm)</pre>
> # the target parameter are the coefficient values that minimize the L2 risk function
> workMSM
Call: glm(formula = Ya ~ A + V + A * V, data = X.msm)
Coefficients:
(Intercept)
                                             A:V
    0.1442 2.9954 0.2987 -0.3125
Degrees of Freedom: 9999 Total (i.e. Null); 9996 Residual
                15050
Null Deviance:
Residual Deviance: 5661
                         AIC: 22700
> # creating a plot
> X.msm1 <- X.msm[X.msm$A==1,]
> X.msm0 <- X.msm[X.msm$A==0,]
> # getting predicted outcomes for each unit given the txt & control
> E.Y1V <- predict(workMSM, newdata=X.msm1, type='response')
> E.YOV <- predict(workMSM, newdata=X.msm0, type='response')
```

Projecting the true causal curve onto the working marginal structural model yields coefficients of  $\beta_0^* = 0.14$ ,  $\beta_1^* = 3$ ,  $\beta_2^* = 0.3$  and  $\beta_3^* = -0.31$ . These coefficients summarize how the counterfactual mean outcome changes a function of treatment A and age V. Specifically, the coefficient on the interaction between the intervention and age is non-zero, indicating that the effect of RUTF on counterfactual mean weight gain is modified by age. See Figure 2.





Solution Fig. 2: Plot of the counterfactual outcomes  $Y_a$  as a function of age (V) and treatment group (a). The predicted counterfactual outcomes, according to the working model, are shown as lines: blue for RUTF and red for the standard supplement.

### 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.