R Assignment 1

PH252D Fall 2013
Introduction to Causal Inference

Assigned: October 14, 2013
Due: October 21, 2013
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

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:

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

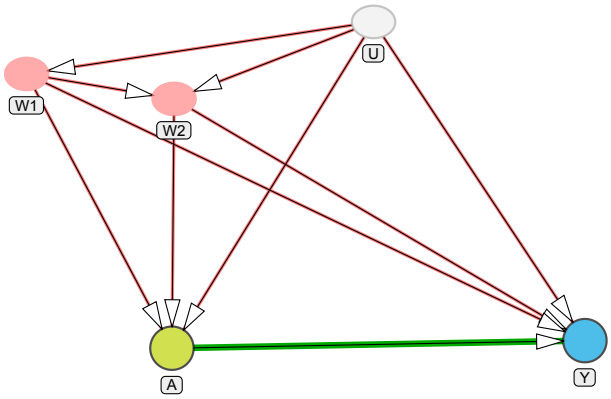
1. **Draw the accompanying DAG.**
2. **Are there any exclusion restrictions?**
3. **Are there any independence assumptions?**
4. **Define the counterfactual outcomes of interest with formal notation and in words. How are counterfactuals derived?**
5. **Suppose we are interested in the average treatment 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?**

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. See Solution Fig. 1.

2. 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 $W2$.)

3. There are no independence assumptions explicitly specified in the background above. (If you made any independence assumptions, you should justify them.)

4. 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, he/she received RUTF ($A = 1$). Y_0 is a child's weight gain if, possibly contrary to fact, he/she received the standard supplement ($A = 0$).
   - Counterfactuals are derived by intervening on the SCM to set $A = a$. The distribution of counterfactuals is impled by the joint distribution of (U, X), denoted $P_{U,X}$. The SCM $\mathcal{M}^F$ gives us a model for the set of possible counterfactual distributions.

5. 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:

$$\begin{aligned}\Psi^F(P_{U,X}) &= E_{U,X}(Y_1) - E_{U,X}(Y_0) \\ &= E_{U,X}[f_Y(W1, W2, 1, U_Y)] - E_{U,X}[f_Y(W1, W2, 0, U_Y)]\end{aligned}$$

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

6. We assume the observed data $O = (W1, W2, A, Y)$ were generated by sampling n i.i.d. times from a data generating system contained in $\mathcal{M}^F$. This provides a link between the causal model $\mathcal{M}^F$ and

the observed data O. The distribution of the exogenous variables U and the structural equations F identify the distribution of the endogenous variables X and thus the distribution of the observed data O, which is denoted P_0.
- We have not placed any restrictions on the statistical model $\mathcal{M}$, which is thereby non-parameteric.

7. In the SCM $\mathcal{M}^F$, the target causal quantity is not identified under the backdoor criterion. In other words, we cannot identify $\Psi^F(P_{U,X})$ as a parameter of the observed data distribution P_0 without additional assumptions. A sufficient, but not minimal, identifiability assumption is that all exogenous errors are independent. Other possibilities include $U_A \perp\!\!\!\perp U_Y$ and (i) $U_A \perp\!\!\!\perp U_{W1},\ U_A \perp\!\!\!\perp U_{W2}$ or (ii) $U_Y \perp\!\!\!\perp U_{W1},\ U_Y \perp\!\!\!\perp U_{W2}$.
- We use $\mathcal{M}^{F*}$ to denote the original SCM augmented with additional assumptions needed for identifiability. We introduce this "working" SCM to keep our real knowledge separate from our wished identifiability assumptions. Under $\mathcal{M}^{F*}$, the backdoor criteria holds conditionally on $(W1, W2)$. The backdoor criterion implies that the counterfactual outcome Y_a is conditionally independent from the intervention A, given $W1$ and $W2$.

8. Under the working SCM $\mathcal{M}^{F*}$, the average treatment effect $\Psi^F(P_{U,X})$ is identified using the G-Computation formula:

$$\begin{aligned}\Psi(P_0) &= E_0\big[E_0(Y|A=1,W1,W2) - E_0(Y|A=0,W1,W2)\big] \\ &= \sum_{w1,w2}\big[E_0(Y|A=1,w1,w2) - E_0(Y|A=0,w1,w2)\big]P_0(w1,w2)\end{aligned}$$

The statistical estimand $\Psi(P_0)$ is the difference in the strata-specific conditional mean weight gain under the intervention and control, averaged with respect to the distribution of baseline covariates (water access and disease history).

Formally, the parameter Ψ is a mapping from the statistical model $\mathcal{M}$ to the parameter space $\mathbb{R}$. In other words, it is a function that takes as input any distribution compatible with the statistical model $\mathcal{M}$ and gives as output a value in the paramter space.

9. To identify $E_{U,X}(Y_a)$ using the G-Computation formula, we need the conditional mean outcome $E_0(Y|A,W)$ to be well-defined for all possible values of the exposure and covariates. Specifically, there to be a positive probability of receiving RUTF $(A=1)$ and the standard supplement $(A=0)$ within all possible strata of water access $(W1)$ and disease history $(W2)$:

$$\begin{aligned}&min_{a\in\mathcal{A}}\ P_0(A=a|W1=w1,W2=w2) > 0, \\ &\text{for all } (w1,w2) \text{ for which } P_0(W1=w1,W2=w2) > 0\end{aligned}$$

The positivity assumption is probably reasonable here. In the target population of school-aged children in a resource-limited country, there is a positive probability of receiving RUTF or the standard supplement within all strata of potable water access and disease history.

2 A specific data generating process

The above SCM is compatible with many possible data generating processes. Recall $\mathcal{M}^F$ is a causal model for the set of possible distributions $P_{U,X}$ for (U,X). Now consider a specific data generating process (unknown to the investigators). 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 logistic function.

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

Solution:

1. For this data generating process, the counterfactual mean outcome is

   $$\begin{aligned} E_{U,X}[Y_a] &= E_{U,X}[4^*a + 0.7^*W1 - 2^*a^*W2 + U_Y] \\ &= 4^*a + 0.7^*E_{U,X}[W1] - 2^*a^*E_{U,X}[W2] + E_{U,X}[U_Y] \end{aligned}$$

   The marginal expectation of $W1$ is $E_{U,X}[W1] = 0.20$. The marginal expectation of $W2$ is

   $$\begin{aligned} E_{U,X}[W2] &= E_{U,X}(W2|W1 = 1)P_{U,X}(W1 = 1) + E_{U,X}(W2|W1 = 0)P_{U,X}(W1 = 0) \\ &= expit(0.5^*1) * 0.20 + expit(0.5^*0) * (1 - 0.20) \end{aligned}$$

   The expectation of the exogenous factor U_Y is zero.

Plugging these values into the target causal parameter, we have

$$\begin{aligned}\Psi^F(P_{U,X}) =& E_{U,X}(Y_1) - E_{U,X}(Y_0)\\ =& 4^*1 + 0.7^*E_{U,X}[W1] - 2^*1^*E_{U,X}[W2] + E_{U,X}[U_Y]\\ & - \big(4^*0 + 0.7^*E_{U,X}[W1] - 2^*0^*E_{U,X}[W2] + E_{U,X}[U_Y]\big)\\ =& 4^*1 + 0.7^*0.20 - 2^*1^*\big[expit(0.5^*1) * 0.20 + expit(0.5^*0) * (1 - 0.20)\big] - \big(0.7^*0.20\big)\\ \approx& 3.09 - 0.14\end{aligned}$$

Therefore, the average treatment effect is

$$\Psi^F(P_{U,X}) = E_{U,X}(Y_1) - E_{U,X}(Y_0) \approx 3.09 - 0.14 = 2.95$$

```
> # corresponding R code for E_{U,X}(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

$$\begin{aligned}E_{U,X}(Y_a) &= E_{U,X}\big[E_{U,X}(Y_a|W1, W2)\big]\\ &= \sum_{w1,w2}\Big[E_{U,X}(Y_a|W1 = w1, W2 = w2)P_{U,X}(W1 = w1, W2 = w2)\Big]\\ &= \sum_{w1,w2}\Big[E_{U,X}(Y_a|W1 = w1, W2 = w2)P_{U,X}(W2 = w2|W1 = w1)P_{U,X}(W1 = w1)\Big]\end{aligned}$$

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

$$E_{U,X}(Y_a|W1, W2) = 4^*a + 0.7^*W1 - 2^*a^*W2$$

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

$$\begin{aligned}P_{U,X}(W2 = 1|W1 = w1) &= expit(0.5w1)\\ P_{U,X}(W2 = 0|W1 = w1) &= 1 - expit(0.5w1)\end{aligned}$$

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

$$\begin{aligned}E_{U,X}(Y_1) =& E_{U,X}(Y_1|W1 = 1, W2 = 1)P_{U,X}(W2 = 1|W1 = 1)P_{U,X}(W1 = 1)\\ &+ E_{U,X}(Y_1|W1 = 0, W2 = 1)P_{U,X}(W2 = 1|W1 = 0)P_{U,X}(W1 = 0)\\ &+ E_{U,X}(Y_1|W1 = 1, W2 = 0)P_{U,X}(W2 = 0|W1 = 1)P_{U,X}(W1 = 1)\\ &+ E_{U,X}(Y_1|W1 = 0, W2 = 0)P_{U,X}(W2 = 0|W1 = 0)P_{U,X}(W1 = 0)\\ &= (4 * 1 + 0.7 * 1 - 2 * 1 * 1)expit(0.5 * 1)0.2\\ &+ (4 * 1 + 0.7 * 0 - 2 * 1 * 1)expit(0.5 * 0)0.8\\ &+ (4 * 1 + 0.7 * 1 - 2 * 1 * 0)(1 - expit(0.5 * 1))0.2\\ &+ (4 * 1 + 0.7 * 0 - 2 * 1 * 0)(1 - expit(0.5 * 0))0.8\\ &\approx 3.09\end{aligned}$$

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

$$\begin{aligned}E_{U,X}(Y_0) =& (4 * 0 + 0.7 * 1 - 2 * 0 * 1)expit(0.5 * 1)0.2\\ &+ (4 * 0 + 0.7 * 0 - 2 * 0 * 1)expit(0.5 * 0)0.8\\ &+ (4 * 0 + 0.7 * 1 - 2 * 0 * 0)(1 - expit(0.5 * 1))0.2\\ &+ (4 * 0 + 0.7 * 0 - 2 * 0 * 0)(1 - expit(0.5 * 0))0.8\\ &= 0.14\end{aligned}$$

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

2.1 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 $\Psi^F(P_{U,X})$.**

Solution:

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

> # 2. set the number of draws n=5000
> n = 5000

> # 3. sample the exogenous 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)

> # 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))
> 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.0<- 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         Y.0
1  0  0 0 -0.39069139 3.609309 -0.39069139
2  0  1 0  0.27579209 2.275792  0.27579209
3  0  1 0  0.13800411 2.138004  0.13800411
4  0  0 0 -0.03862696 3.961373 -0.03862696
5  0  1 1  2.08010486 2.080105  0.08010486
6  0  0 0 -0.02693322 3.973067 -0.02693322

> summary(X)

       W1               W2               A                Y
 Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :-0.91451
 1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.: 0.08653
 Median :0.0000   Median :1.0000   Median :1.0000   Median : 1.66352
 Mean   :0.1854   Mean   :0.5184   Mean   :0.5258   Mean   : 1.66258
 3rd Qu.:0.0000   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.: 3.05952
 Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   : 5.32635
      Y.1              Y.0
 Min.   :1.025   Min.   :-0.9749
 1st Qu.:2.090   1st Qu.:-0.1505
 Median :3.032   Median : 0.0816
 Mean   :3.098   Mean   : 0.1346
 3rd Qu.:4.044   3rd Qu.: 0.3800
 Max.   :5.361   Max.   : 1.6252

> # 7. evaluate the causal parameter
> Psi.F<- mean(Y.1) - mean(Y.0)
> Psi.F

[1] 2.9632
```

3 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:

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

Then the target parameter β is the defined as a projection of the true causal curve onto a working model. Specifically, β is the set of coefficient 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.

3.1 A specific data generating process:

Consider a new data generating process (one of many compatible with the above 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.**

Solution:

```
> # set the number of draws n=5000
> n = 5000
> # sample the exogenous 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-intervention covariates and the counterfactuals
> V <- 2 + U.V
> W1 <- as.numeric( U.W1 < 0.2)
> W2 <- as.numeric( U.W2 < plogis(0.5*W1) )
> Y.1 <- 2*1 + 0.3*W1 + 2*1*W2 + 0.5*1*V + U.Y
> Y.0 <- 2*0 + 0.3*W1 + 2*0*W2 + 0.5*0*V + U.Y
> mean(Y.1)

[1] 4.882042

> mean(Y.0)

[1] 0.06295943

> # 2. create data frame X.msm
> V.msm<- rep(V,2)
> A.msm<- c( rep(1,n), rep(0,n))
> Y.a.msm<- c(Y.1, Y.0)
> X.msm<- data.frame(V.msm, A.msm, Y.a.msm)
> head(X.msm)

     V.msm A.msm  Y.a.msm
1 2.865001     1 5.636531
2 2.561970     1 5.490133
3 3.898006     1 3.927301
4 3.093666     1 3.608055
5 2.072531     1 3.363286
6 4.267243     1 6.162491

> tail(X.msm)

         V.msm A.msm     Y.a.msm
9995  3.540554     0 -0.06901472
9996  4.697067     0  0.02094975
9997  4.275578     0  0.06574547
9998  2.540256     0 -0.01601664
9999  2.401097     0  0.02757479
10000 2.411879     0  0.06608637
```

```
> # 3. evaluate the causal parameter
> #We can use ordinary least squares regression to obtain betas.
> workMSM<- glm(Y.a.msm ~ A.msm*V.msm, data=X.msm)
> #
> # the target parameter are the coefficient values that minimize the L2 risk function
> workMSM

Call:  glm(formula = Y.a.msm ~ A.msm * V.msm, data = X.msm)

Coefficients:
(Intercept)        A.msm        V.msm  A.msm:V.msm
   0.069793     3.082615    -0.001946     0.494584

Degrees of Freedom: 9999 Total (i.e. Null);  9996 Residual
Null Deviance:            64320
Residual Deviance: 5349         AIC: 22130
```

Projecting the true causal curve onto the working marginal structural model yields coefficients of $\beta_0 = 0.07$, $\beta_1 = 3$, $\beta_2 = 0$ and $\beta_3 = 0.5$. 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.

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.