R Assignment 1: Causal Parameters & Simulations in R

Alvaro J. Castro Rivadeneira

September 30, 2021

```
## Loading required package: ggplot2

##
## Attaching package: 'ggdag'

## The following object is masked from 'package:stats':
##
##     filter
```

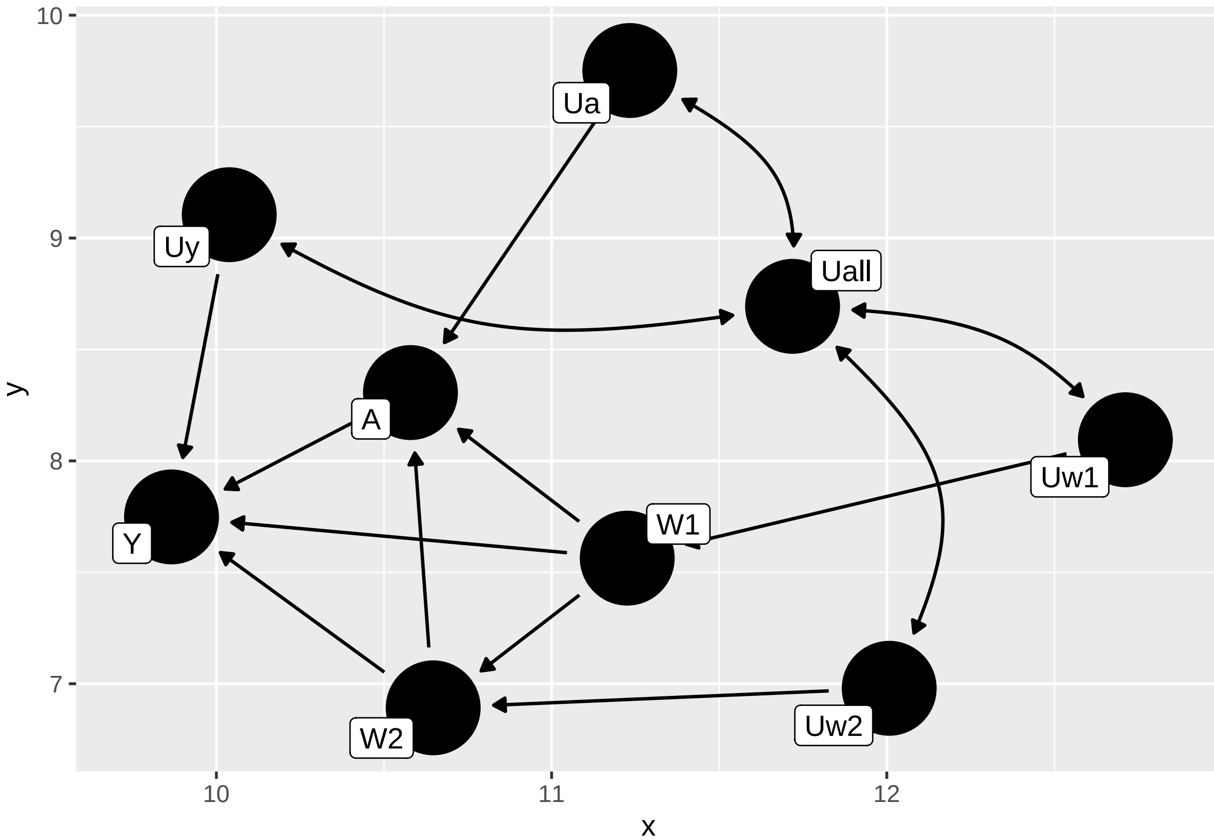
2 Steps 1-2 of the Roadmap

Step 1: Causal model representing real knowledge

(a) Draw the accompanying directed acyclic graph (DAG).

```
rutf_wt_gain <- dagify(weight_gain ~ RUTF + potable_h20 + inf_dis + Uy,
                       RUTF ~ potable_h20 + inf_dis + Ua,
                       inf_dis ~ potable_h20 + Uw2,
                       potable_h20 ~ Uw1,
                       Uall ~~ Uy + Ua + Uw2 + Uw1,
                       labels = c("weight_gain" = "Y",
                                  "RUTF" = "A",
                                  "potable_h20" = "W1",
                                  "inf_dis" = "W2",
                                  "Uy" = "Uy",
                                  "Ua" = "Ua",
                                  "Uw1" = "Uw1",
                                  "Uw2" = "Uw2",
                                  "Uall" = "Uall"),
                       exposure = "RUTF",
                       outcome = "weight_gain")

ggdag(rutf_wt_gain, text = FALSE, use_labels = "label")
```



(b) Are there any exclusion restrictions? Recall we are working with recursive (time-ordered) structural causal models.

We do not have any exclusion restrictions, because we are not leaving variables out of a parent set of the exposure. In our SCM, all endogenous variables have an effect on the outcome.

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

There are no justified restrictions on the set of allowed distributions for the background variables U, which is expressed through their relationship via Uall.

Step 2: Counterfactuals & causal parameter

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

The counterffactuals of interest are $(Y_a : a \in A = \{0, 1\})$, where Y_1 is the counterfactual weight gain if, possibly contrary to fact, school aged children received the RUTF supplement; and Y_0 is the counterfactual weight gain if, possibly contrary to fact, school aged children did not receive the RUTF supplement.

(b) How are counterfactuals derived?

Counterfactuals are derived by setting the exposure (or treatment) A to a given value a, so that A = a.

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

The target causal parameter is the average treatment effect of RUTF supplements.

$$\Psi^*(\mathbb{P}^*) = \mathbb{E}^*(Y_1) - \mathbb{E}^*(Y_0)$$
$$= \mathbb{E}^*[f_Y(W1, W2, 1, U_Y)] - \mathbb{E}^*[f_Y(W1, W2, 0, U_Y)]$$

This is the difference in the expected counterfactual weight gain if all school-aged children were given RUTF supplements and the expected counterfactual weight gain if all school-aged children were not given RUTF supplements.

3 A specific data generating process

3.1 Closed form evaluation of the target causal parameter

Evaluate the target causal parameter $\Psi^*(\mathbf{P}^*)$ in closed form (i.e., by hand) for this data generating process.

$\Psi^*(\mathbf{P}^*) = \mathbb{E}^*(Y_1 - Y_0)\mathbb{E}^*[Y_a] = \mathbb{E}^*[4*A + 0.7*W1 - 2*A*W2 + U_Y]\mathbb{E}^*(Y_1) = (4*1 + 0.7*0.2 - 2*1*logit^{-1}(0.5*0.2)) + \mathbb{E}^*[U_Y] = (4.$

```
Ey1=4.14-(2*(exp(0.1)/(1+(exp(0.1)))))
Ey1
```

```
## [1] 3.090042
```

$\mathbb{E}^*(Y_0) = (4*0 + 0.7*0.2 - 2*0*logit^{-1}(0.5*0.2)) + \mathbb{E}^*[U_Y] = 0.14 + \mathbb{E}^*[U_Y]\Psi^*(\mathbf{P}^*) = \mathbb{E}^*(Y_1 - Y_0) = 3.09 - 0.14\Psi^*(\mathbf{P}^*) = 2.95$

3.2 Translating this data generating process for $\mathbb{P}^*$ into simulations, generating counterfactual outcomes and evaluating the target causal parameter.

1. First set the seed to 252.

```
set.seed(252)
```

2. Set `n=50,000` as the number of i.i.d. draws from the data generating process.

```
n <- 50000
```

3. Simulate the background factors $\boldsymbol{U}$. Note the syntax for `rnorm`.

```
U.W1 <- runif(n, min=0, max=1)
U.W2 <- runif(n, min=0, max=1)
U.A <- runif(n, min=0, max=1)
U.Y <- rnorm(n, mean=0, sd=0.3)
```

4. Evaluate the structural equations $\mathcal{F}$ to deterministically generate the endogenous nodes X. Recall that the $logit^{-1}$ function is given by the `plogis` function in `R`.

```r
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

X <- data.frame(W1, W2, A, Y)
head(X)
```

```
##   W1 W2 A            Y
## 1  0  0 0 -0.188412377
## 2  0  0 0  0.007554149
## 3  0  1 0  0.485863200
## 4  0  0 0  0.207939977
## 5  0  1 0 -0.189080671
## 6  0  1 0 -0.017563026
```

```r
tail(X)
```

```
##       W1 W2 A         Y
## 49995  1  1 0 1.3625401
## 49996  1  1 0 0.8510302
## 49997  1  1 1 2.1844276
## 49998  0  1 1 1.8360129
## 49999  0  0 1 3.6332872
## 50000  0  1 1 1.6108151
```

```r
summary(X)
```

```
##        W1               W2               A                Y
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :-1.15726
##  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
##  Mean   :0.1976   Mean   :0.5254   Mean   :0.5292   Mean   : 1.67314
##  3rd Qu.:0.0000   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.: 3.04173
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   : 5.66234
```

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.

```r
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 a data frame `X` to hold the values of the endogenous factors (W_1, W_2, 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 $A = a$?

```r
X <- data.frame(X, Y.1, Y.0, U.Y)
head(X)
```

```
##    W1  W2 A            Y      Y.1          Y.0          U.Y
## 1  0   0 0 -0.188412377 3.811588 -0.188412377 -0.188412377
## 2  0   0 0  0.007554149 4.007554  0.007554149  0.007554149
## 3  0   1 0  0.485863200 2.485863  0.485863200  0.485863200
## 4  0   0 0  0.207939977 4.207940  0.207939977  0.207939977
## 5  0   1 0 -0.189080671 1.810919 -0.189080671 -0.189080671
## 6  0   1 0 -0.017563026 1.982437 -0.017563026 -0.017563026
```

```r
tail(X)
```

```
##        W1  W2 A         Y      Y.1        Y.0        U.Y
## 49995  1   1 0 1.3625401 3.362540  1.3625401  0.6625401
## 49996  1   1 0 0.8510302 2.851030  0.8510302  0.1510302
## 49997  1   1 1 2.1844276 2.184428  0.1844276 -0.5155724
## 49998  0   1 1 1.8360129 1.836013 -0.1639871 -0.1639871
## 49999  0   0 1 3.6332872 3.633287 -0.3667128 -0.3667128
## 50000  0   1 1 1.6108151 1.610815 -0.3891849 -0.3891849
```

```r
summary(X)
```

```
##        W1               W2               A                Y
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :-1.15726
##  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
##  Mean   :0.1976   Mean   :0.5254   Mean   :0.5292   Mean   : 1.67314
##  3rd Qu.:0.0000   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.: 3.04173
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   : 5.66234
##       Y.1              Y.0                U.Y
##  Min.   :0.8427   Min.   :-1.15726   Min.   :-1.1572597
##  1st Qu.:2.0867   1st Qu.:-0.15171   1st Qu.:-0.2068829
##  Median :2.9530   Median : 0.08838   Median :-0.0000859
##  Mean   :3.0871   Mean   : 0.13794   Mean   :-0.0003948
##  3rd Qu.:4.0423   3rd Qu.: 0.38180   3rd Qu.: 0.2053934
##  Max.   :5.6670   Max.   : 1.94849   Max.   : 1.2484928
```

To evaluate whether the counterfactual value Y_a equal the observed Y when $A = a$, we can do the following:

```r
Ya1 <- mean (Y.1 + U.Y)
Ya1
```

```
## [1] 3.086664
```

As we can see, this is very close to our prediction of 3.09, and so we can say that the counterfactual value Y_a equals the observed Y when $A = a$.

7. Using these simulations, evaluate the causal parameter $\Psi^*(\mathbb{P}^*)$ for this population of 50,000 units.

```r
Psi.star<- mean(Y.1 - Y.0)
Psi.star
```

```
## [1] 2.94912
```

This is almost identical to our previously calculated value of 2.95, and thus, they match.

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

4 Defining the target causal parameter with a working MSM

4.1 A specific data generating process

1. For $n = 5,000$ children, generate the background 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.

```
set.seed(252)
n <- 5000

U.V <- runif(n, min=0, max=3)
U.W1 <- runif(n, min=0, max=1)
U.W2 <- runif(n, min=0, max=1)
U.A <- runif(n, min=0, max=1)
U.Y <- rnorm(n, mean=0, sd=0.1)

V = 2 + U.V
W1 <- as.numeric(U.W1 < 0.2)
W2 <- as.numeric(U.W2 < plogis(0.5*W1))
A <- as.numeric(U.A < plogis(W1*W2 + (V/5)))
Y <- 4*A + 0.7*W1 - 2*A*W2 + 0.3*V - 0.3*A*V + U.Y

X <- data.frame(V, W1, W2, A, Y)

Y.1 <- 4*1 + 0.7*W1 - 2*1*W2 + 0.3*V - 0.3*1*V + U.Y
Y.0 <- 4*0 + 0.7*W1 - 2*0*W2 + 0.3*V - 0.3*0*V + U.Y

X <- data.frame(X, Y.1, Y.0, U.Y)
head(X)
```

```
##          V W1 W2 A         Y      Y.1       Y.0         U.Y
## 1 4.692824  0  0 1 3.9520362 3.952036 1.3598833 -0.04796384
## 2 4.136408  1  0 1 4.6597935 4.659794 1.9007158 -0.04020647
## 3 2.982279  0  0 0 0.6795578 3.784874 0.6795578 -0.21512580
## 4 4.303568  0  0 1 4.0158098 4.015810 1.3068801  0.01580981
## 5 4.049335  0  1 1 1.8110843 1.811084 1.0258849 -0.18891568
## 6 3.114801  0  0 0 0.9810607 4.046621 0.9810607  0.04662055
```

```
tail(X)
```

```
##             V W1 W2 A         Y      Y.1       Y.0          U.Y
## 4995 2.581204  0  1 0 0.7393272 1.964966 0.7393272 -0.035034027
## 4996 4.108090  0  0 0 1.1830232 3.950596 1.1830232 -0.049403753
## 4997 4.641096  1  1 1 2.6624310 2.662431 2.0547598 -0.037569016
## 4998 3.756570  0  0 1 4.0033891 4.003389 1.1303600  0.003389131
## 4999 4.970072  1  1 1 2.7701569 2.770157 2.2611784  0.070156886
## 5000 3.158294  0  1 1 1.8890925 1.889093 0.8365807 -0.110907477
```

```
summary(X)
```

```
##        V               W1               W2                A        
##  Min.   :2.001   Min.   :0.000   Min.   :0.0000   Min.   :0.000  
##  1st Qu.:2.788   1st Qu.:0.000   1st Qu.:0.0000   1st Qu.:0.000  
##  Median :3.550   Median :0.000   Median :1.0000   Median :1.000  
##  Mean   :3.527   Mean   :0.205   Mean   :0.5266   Mean   :0.689  
##  3rd Qu.:4.270   3rd Qu.:0.000   3rd Qu.:1.0000   3rd Qu.:1.000  
##  Max.   :5.000   Max.   :1.000   Max.   :1.0000   Max.   :1.000  
##        Y               Y.1              Y.0               U.Y            
##  Min.   :0.3841   Min.   :1.581   Min.   :0.3773   Min.   :-0.4188832  
##  1st Qu.:1.3959   1st Qu.:2.035   1st Qu.:0.8980   1st Qu.:-0.0658665  
##  Median :2.0639   Median :2.789   Median :1.1791   Median : 0.0009475  
##  Mean   :2.4662   Mean   :3.092   Mean   :1.2035   Mean   : 0.0018408  
##  3rd Qu.:3.9305   3rd Qu.:4.015   3rd Qu.:1.4292   3rd Qu.: 0.0708916  
##  Max.   :4.9522   Max.   :4.952   Max.   :2.3945   Max.   : 0.3440848
```

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

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

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.

```
X.msm <- data.frame(V, A, Y)
head(X.msm)
```

```
##          V A         Y
## 1 4.692824 1 3.9520362
## 2 4.136408 1 4.6597935
## 3 2.982279 0 0.6795578
## 4 4.303568 1 4.0158098
## 5 4.049335 1 1.8110843
## 6 3.114801 0 0.9810607
```

3. Evaluate the target causal parameter. We have defined the target parameter using the last 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`.

```
X.msm <- glm(Y ~ A*V)
X.msm
```

```
## 
## Call:  glm(formula = Y ~ A * V)
```

```
##
## Coefficients:
## (Intercept)            A            V          A:V
##      0.1280       2.9146       0.2929      -0.2842
##
## Degrees of Freedom: 4999 Total (i.e. Null);  4996 Residual
## Null Deviance:       7755
## Residual Deviance: 3567  AIC: 12510
```

4. Interpret the results.

Broadly, in this linear model, the intervention has a large positive effect on weight gain, whose effect decreases as age increases due to the effect of the interaction between these two.

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

```
# Optional: bonus section
```