Bob Bell

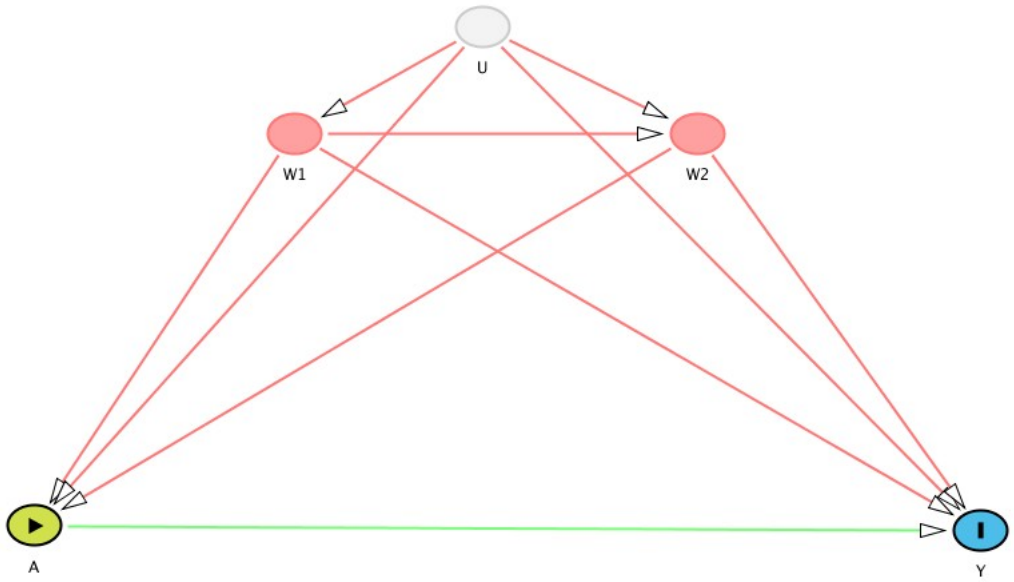
10.21.2013

PH252D

R Assignment 1

1 Background

1. Directed Acyclic Graph:



2. There are no exclusion restrictions because the structural equation for every endogenous variable includes in its parent set all endogenous variables before it (with respect to time ordering).

3. There are no independence assumptions, as we have not assumed any exogenous/background (U) variables are independent from other U's.

4. Counterfactual outcome of interest are$(Y_a : a \in \mathcal{A} = \{0, 1\})$where $Y_a = f_Y(W, a, U_Y)$.$Y_0$is the counterfactual child's weight gain in pounds at the study termination if the child had received the standard supplement.$Y_1$is the counterfactual child's weight gain in pounds at the study termination if the child received RUTF.

5. The target causal parameter is the average causal effect given by the following notation:

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

Average treatment effect is the difference in the expected counterfactual weight gain of all children at the end of the study if all children received RUTF and the expected value of the counterfactual weight gain of all children at the end of the study if all children received the standard supplement.

6. Observed data is generated by the data-generating system contained by our SCM. Therefore, $O \subseteq X$. The distribution of the background variables P_U and the structural equations$F$ identify the distribution of$X$ and thus the distribution for$O$. Our observed data was generated by sampling n i.i.d. times from the data generating system specified in our causal model, giving us $O_1, O_2, \ldots, O_n$from the probability distributionP_0. The statistical model has no exclusion restrictions or independence assumptions. Thus, it is non-parametric. This true, unknown distribution of the observed data and the statistical model can be translated into the structural causal model$\mathcal{M}^F$.

7. Using the backdoor criterion, the target causal parameter is not identified. I need to make some assumptions to establish identifiability. If $U_A \perp U_Y$, then we can establish identifiability controlling for $\{W_1, W_2\}$. This original SCM with additional assumptions is denoted$\mathcal{M}^{F*}$.

8. Target parameter (statistical estimand) is:

$$\Psi(P_0) = E_0[E_0(Y \mid A = 1, W1, W2) - E_0(Y \mid A = 0, W1, W2)]$$

$$= \sum_{w1,w2} [E_0(Y \mid A = 1, W1 = w1, W2 = w2) - E_0(Y \mid A = 0, W1 = w1, W2 = w2)] P_0(W1 = w1, W2 = w2)$$

9. Positivity Assumption:

$$\min_{a \in A} P_0(A = a \mid W1 = w1, W2 = w2) > 0 \, (i.e.) \, 0 < P_0(A = 1 \mid W1 = w1, W2 = w2) < 1$$

for all $w1, w2$ for which $P_0(W1 = w1, W2 = w2) > 0$. I think the positivity assumption is reasonable here. Theoretically, for all covariate pairs, it is possible to observe children receiving both the RUTF and standard supplements. Practically, I could also make this assumption as we only have 4 covariate pairs.

2 A specific data generating process

1. Assemble components of target causal parameter and estimate:

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

$P_{U,X}(W1 = 0) = 0.80$

$E_{U,X}(W2 \mid W1 = 1) = expit(0.5 * 1) = \frac{1}{1+e^{-0.5}} \approx 0.62246$

$E_{U,X}(W2 \mid W1 = 0) = expit(0.5 * 0) = \frac{1}{1+e^{-0}} \approx 0.50000$

$E_{U,X}(W2) = 0.62246 * 0.20 + 0.5 * 0.8 \approx 0.52449$

$E_{U,X}(Y_1) - E_{U,X}(Y_0) = E_{U,X}[4 * 1] + E_{U,X}[0.7 * W1] - E_{U,X}[2 * 1 * W2] + E_{U,X}[U_Y] - E_{U,X}[4 * 0] - E_{U,X}[0.7 * W1] + E_{U,X}[2 * 0 * W2] - E_{U,X}[U_Y]$

$= E_{U,X}[4 * 1] - E_{U,X}[2 * 1 * W2]$

$= 4 - 2 * E_{U,X}[W2]$

$=4 - 2 * 0.52449$

$= 2.9510$

2. Interpret this estimand: On average, for all children, the impact of the RUTF increases the child's weight gain at the end of the study by 2.9510 pounds.

2.1 Translating data generating process into simulations.

```
# 1. setting the seed
set.seed(252)

#2. setting the number of observations
n<- 5000

#3. simulating the Us
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. Given the random input, deterministically evaluate the F equations.
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. intervene to set A=a and generate the counterfactual outcomes Y.a
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 with endogenous factors and counterfactual outcomes
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                 Y.1
 Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :-0.91451   Min.   :1.025
 1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.: 0.08653   1st Qu.:2.090
 Median :0.0000   Median :1.0000   Median :1.0000   Median : 1.66352   Median :3.032
 Mean   :0.1854   Mean   :0.5184   Mean   :0.5258   Mean   : 1.66258   Mean   :3.098
 3rd Qu.:0.0000   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.: 3.05952   3rd Qu.:4.044
 Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   : 5.32635   Max.   :5.361
      Y.0
 Min.   :-0.9749
 1st Qu.:-0.1505
 Median : 0.0816
 Mean   : 0.1346
 3rd Qu.: 0.3800
 Max.   : 1.6252

# 7. Evaluate the causal parameter
Psi.F<- mean(Y.1 - Y.0)
Psi.F
[1] 2.9632
```

3 Defining target causal parameter with working MSM

```
#1 Generate exogenous factors, covariates, and counterfactuals
U.V.msm <- runif(n, min=0, max=3)
U.W1.msm <- U.W1
U.W2.msm <- U.W2
U.A.msm <- U.A
U.Y.msm <- rnorm(n, mean=0, sd=0.1)
V.msm <- 2 + U.V.msm
W1.msm <- W1
W2.msm <- W2
A.msm <- as.numeric(U.A.msm<plogis(W1.msm*W2.msm + V.msm*0.2))
Y.msm <- 2*A.msm + 0.3*W1.msm + 2*A.msm*W2.msm + 0.5*A.msm*V.msm + U.Y.msm
Y.1.msm <- 2*1 + 0.3*W1.msm + 2*1*W2.msm + 0.5*1*V.msm + U.Y.msm
```

```
Y.0.msm <- 2*0 + 0.3*w1.msm + 2*0*w2.msm + 0.5*0*V.msm + U.Y.msm
Y.a.msm <- c(Y.1.msm, Y.0.msm)
a.msm<- c( rep(1,n), rep(0, n) )

#2 Create data frame X.msm
X.msm<- data.frame(V.msm, a.msm, Y.a.msm)

#3 Evaluate target causal parameter
workMSM <- glm(formula=Y.a.msm ~ a.msm*V.msm, data=X.msm)
workMSM

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

Coefficients:
(Intercept)        a.msm        V.msm  a.msm:V.msm
  0.0552329    3.0900971   -0.0002508    0.4848198

Degrees of Freedom: 9999 Total (i.e. Null);  9996 Residual
Null Deviance:       63640
Residual Deviance: 5348   AIC: 22130
```

4. Interpret results: Treatment A increases outcome Y by approximately 3 units (or 3 pounds at the end of the study). This is somewhat close to the estimands we calculated earlier. V (age) has a relatively negligible effect on counterfactual outcome. The effect term (age*treatment) increases the child's weight by approximately ½ pound by the end of the study.