R Assignment 4

PH252D Fall 2013
Introduction to Causal Inference

Assigned: November 13, 2013
Due: November 27, 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.

1 Background and Causal Road Map

"Russian Health Campaign Allows Train Users In Moscow To Pay In Squats"
The Huffington Post UK
"Want a free journey on the Tube in Moscow? Drop down and give 30 squats. In an effort to promote the upcoming Winter Olympic Games in Sochi, Moscow city officials and the Russian Olympic Committee are allowing subway riders to sweat it out to get to work. Instead of paying the regular 30 rubles (57p), commuters can now perform 30 squats at Vystavochnaya station..."
http://www.huffingtonpost.co.uk/2013/11/12/russia-moscow-train-squats_n_4260746.html

Consider a hypothetical intervention on the BART system, where riders will be given discounted tickets based on the number burpees they can properly complete (http://en.wikipedia.org/wiki/Burpee_(exercise)). For simplicity assume the minimum is 1 burpee and maximum is 7 burpees. The goal is to estimate the effect of burpees performed on rider's happiness, which is measured on a validated scale from 30-45. We have data on the the following variables

- $W1$: lifestyle with 1 as very active, 2 as somewhat active and 3 as sedentary
- $W2$: gender with 1 as female and 0 as male
- A: number of burpees completed with 1 as the minimum and 7 as the maximum
- Y: happiness which is a continuous scale from 30-45

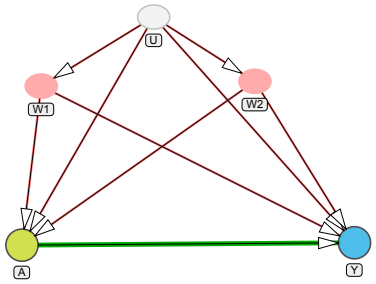

Figure 1: DAG corresponding to the study of burpees and happiness among BART riders.

Causal Roadmap Rundown

This is a very, very quick summary for review. Each step of the road map requires careful thought and consideration.

1. **Specify the Question:**
   What is the causal effect of burpees completed on happiness among BART riders?

2. **Specify the causal model $\mathcal{M}^F$:**
   - Endogenous nodes: $X = (W1, W2, A, Y)$, where $W1$ is lifestyle, $W2$ is gender, A is number of burpees completed and Y is measured happiness.
   - Exogenous nodes: $U = (U_{W1}, U_{W2}, U_A, U_Y) \sim P_U$. There are no independence assumptions.
   - Structural equations F:

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

   Exclusion restrictions: we are assuming that the baseline covariates do not affect each other. We have not specified any functional forms.

3. **Specify the causal parameter:**
   The exposure (burpees completed) has seven possible levels:

$$\mathcal{A} = \{1, 2, 3, 4, 5, 6, 7\}$$

   Therefore, we could consider defining the target causal parameter in terms of all pairwise differences of interest:

$$\Psi^F(P_{U,X}) = E_{U,X}(Y_a) - E_{U,X}(Y_{a'})$$

   where Y_a is the counterfactual outcome (happiness), if possibly contrary to fact, the rider completed $A = a$ burpees. Here, $\Psi^F(P_{U,X})$ is the difference in the expected counterfactual happiness if all riders completed $A = a$ burpees versus if they all riders completed $A = a'$ burpees.

   We could also consider a marginal structural model (MSM) to summarize how the counterfactual mean happiness (Y_a) changes as a function burpees completed a. Consider, for example,

$$E_{U,X}(Y_a) = m(a|\beta) = \beta_0 + \beta_1 a$$

   with $a \in \{1, 2, 3, 4, 5, 6, 7\}$. This specification assumes a linear change in the expected counterfactual happiness with burpees completed. Alternatively, we can consider this a working MSM and define the parameter as the projection of the true causal curve onto this linear model:

$$\beta(P_{U,X}|m) = argmin_{\beta'} \; E_{P_{U,X}} \left[\sum_{a \in \mathcal{A}} (Y_a - m(a|\beta'))^2 g^*(a) \right]$$

   where $g^*(a)$ specifies how much weight we put on specific values of a. For a non-conditional MSM as above, the typical choice for $g^*(a)$ is the marginal probability $P_0(A = a)$.

4. **Specify the link between the SCM and the observed data:**
   The observed data were generated by sampling n independent times from a data generating system described by the structural causal model $\mathcal{M}^F$. This yields n i.i.d. copies of random variable $O = (W1, W2, A, Y) \sim P_0$. The statistical model $\mathcal{M}$ for the set of allowed observed data distributions is non-parametric.

5. **Assess identifiability:**
   In the original SCM $\mathcal{M}^F$, the target causal parameter is not identified from the observed data distribution. A sufficient, but not minimal, assumption is that all of the unmeasured factors 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 M^{F*} to denote the original SCM augmented by the assumptions needed for identifiability. Under $\mathcal{M}^{F*}$, the backdoor criteria will hold conditionally on the set of baseline covariates $W = (W1, W2)$.

To identify $E_{U,X}(Y_a)$ with the G-Computation formula, we also need the positivity assumption to hold

$$min_{a \in \mathcal{A}} \; P_0(A = a|W = w) > 0$$

for all w for which $P_0(W = w) > 0$. In terms of our example, there must be a positive probability of each exposure (number of completed burpees) within strata of lifestyle and gender. As detailed below, the positivity assumption needed for MSM parameters depends on the choice of the numerator $g^*(A)$.

6. **Specify the statistical estimand:**
Under the working SCM $\mathcal{M}^{F*}$, the average treatment effect $\Psi^F(P_{U,X})$ can be identified with the G-Computation formula:

$$\Psi(P_0) = E_{0,W}\big[E_0(Y|A = a, W) - E_0(Y|A = a', W)\big]$$

where $W = (W1, W2)$ is the vector of baseline covariates.

For the target causal parameter defined with an MSM, the statistical parameter is the projection of the conditional mean outcome $E_0(Y|A = a, W)$ onto the MSM $m(a|\beta)$.

2 A specific data generating process

The following code was used to generate the data set `Rassign4.Fa2013.csv`. In this data generating process (one of many compatible with the above SCM), all exogenous errors are independent.

```
> # ----------------
> # generateData - function to generate the observed data + counterfactuals
> # ----------------
> generateData<- function(n){
+
+   W1<- as.integer(runif(n, 1,4)  ) # lifestyle 1,2,3
+   W2<- rbinom(n, size=1, prob= runif(n, 0.02, 0.7)) # gender
+   A<- 1+ rbinom(n, size=6, prob=plogis(0.35 -0.3*W1 +0.5*(1-W2) )) #burpees
+   U.Y<- rnorm(n, 0, sd=0.01)
+   Y<- 30 +1.5*W1 +3*log(A)+.3*(1-W2)*A + U.Y # happiness
+
+   #  the counterfactuals
+   Y.1<- 30 +1.5*W1 +3*log(1)+.3*(1-W2)*1 + U.Y #
+   Y.2<- 30 +1.5*W1 +3*log(2)+.3*(1-W2)*2 + U.Y #
+   Y.3<- 30 +1.5*W1 +3*log(3)+.3*(1-W2)*3 + U.Y #
+   Y.4<- 30 +1.5*W1 +3*log(4)+.3*(1-W2)*4 + U.Y #
+   Y.5<- 30 +1.5*W1 +3*log(5)+.3*(1-W2)*5 + U.Y #
+   Y.6<- 30 +1.5*W1 +3*log(6)+.3*(1-W2)*6 + U.Y #
+   Y.7<- 30 +1.5*W1 +3*log(7)+.3*(1-W2)*7 + U.Y #
+
+   data.frame(W1,W2,A,Y,Y.1, Y.2, Y.3, Y.4,Y.5, Y.6, Y.7)
+ }
> # ----------------

> # observed data
> set.seed(252)
> FullData<- generateData(n=5000)
> ObsData<- subset(FullData, select=c(W1,W2,A,Y))
> write.csv(ObsData, file="Rassign4.Fa2013.csv", row.names=F)
```

1. **Given this specific data generating process, calculate the expected counterfactual outcome $E_{U,X}(Y_a)$, under each exposure level $a \in \mathcal{A}$.**

3 Import and explore data set `Rassign4.Fa2013.csv`.

1. Import the data set and assign it to object `ObsData`.
2. Assign the number of riders to `n`.
3. Use the `summary` function to explore the data.
4. Are there certain covariate combinations with limited variability in the exposure (burpees completed)?
   (a) Create a dummy variable `strataW1W2` for the different combinations of baseline covariates (lifestyle $W1$ and gender $W2$):

```
> strataW1W2<- rep(NA, n)
> strataW1W2[ ObsData$W1==1 & ObsData$W2==1] <- 11
> strataW1W2[ ObsData$W1==2 & ObsData$W2==1] <- 21
> strataW1W2[ ObsData$W1==3 & ObsData$W2==1] <- 31
> strataW1W2[ ObsData$W1==1 & ObsData$W2==0] <- 10
> strataW1W2[ ObsData$W1==2 & ObsData$W2==0] <- 20
> strataW1W2[ ObsData$W1==3 & ObsData$W2==0] <- 30
> # telling R that these are factors
> strataW1W2<- as.factor(strataW1W2)
```

   (b) Use the table function to check the number of riders in each exposure-covariate category. *Note: We are just counting the number of observations in a single sample of size n - not formally evaluating the positivity assumption, which is an assumption on the true data generating process.*
   (c) Comment.

4 IPTW for the statistical estimand equal to the ATE under $\mathcal{M}^{F*}$

Suppose we are interested in the difference in the expected happiness if all riders completed 7 burpees ($A = 7$) and if all riders only completed 1 burpee ($A = 1$):

$$\Psi^F(P_{U,X}) = E_{U,X}(Y_7) - E_{U,X}(Y_1)$$

1. We need to estimate the treatment mechanism $P_0(A|W) = g_0(A|W)$, which is the conditional probability of completing A burpees, given the rider's characteristics. Implement the following code to estimate the treatment mechanism with multinomial logistic regression. You will need the `nnet` package:

```
> library("nnet")
> gAW.reg<-multinom(A~ W1+W2, data=ObsData)
```

2. Predict each rider's probability of his/her observed exposure (burpees completed), given his/her covariates $g_n(A_i|W_i)$:
   (a) Use the `predict` function to obtain the predicted probability of each exposure level, given the rider's covariates. Be sure to specify `type="probs"`.

```
> gAW.pred<- predict(gAW.reg, type="probs")
```

   (b) Create an empty vector `gAW` of length n for the predicted probabilities.
   (c) Among riders with exposure level $A = 1$, assign the appropriate predicted probability:

```
> gAW[ObsData$A==1] <- gAW.pred[ObsData$A==1, "1"]
```

(d) Implement the analogous code for exposure levels $A = 2, \ldots A = 7$:

(e) Use the `summary` function to examine the distribution of predicted probabilities. Any cause for concern?

3. Create the vector `wt` as the inverse of the predicted probabilities. Use the `summary` function to examine the distribution of weights.

4. Evaluate the IPTW estimand:

$$\hat{\Psi}_{IPTW}(P_n) = \frac{1}{n}\sum_{i=1}^{n}\frac{\mathbb{I}(A_i = 7)}{g_n(A_i|W_i)}Y_i - \frac{1}{n}\sum_{i=1}^{n}\frac{\mathbb{I}(A_i = 1)}{g_n(A_i|W_i)}Y_i$$

The first quantity is the weighted mean outcome, where riders completing $A_i = 7$ burpees receive weight $1/g_n(A_i = 7|W_i)$ and riders completing $A_i \neq 7$ burpees receive weight 0. The second quantity is the weighted mean outcome, where riders completing $A_i = 1$ burpees receive weight $1/g_n(A_i = 1|W_i)$ and riders completing $A_i \neq 1$ burpees receive weight 0.

5. Implement the stabilized IPTW estimator (a.k.a. the modified Horvitz-Thompson estimator):

$$\hat{\Psi}_{St.IPTW} = \frac{\frac{1}{n}\sum_{i=1}^{n}\frac{\mathbb{I}(A_i=7)}{g_n(A_i|W_i)}Y_i}{\frac{1}{n}\sum_{i=1}^{n}\frac{\mathbb{I}(A_i=7)}{g_n(A_i|W_i)}} - \frac{\frac{1}{n}\sum_{i=1}^{n}\frac{\mathbb{I}(A_i=1)}{g_n(A_i|W_i)}Y_i}{\frac{1}{n}\sum_{i=1}^{n}\frac{\mathbb{I}(A_i=1)}{g_n(A_i|W_i)}}$$

6. Comment on the results.

5 IPTW & Marginal Structural Models

In the previous section, we focused on the effect of completing the highest level of burpees ($A = 7$) and the lowest level of burpees ($A = 0$) on happiness. Now suppose we want to smooth over categories of burpees. Consider the following MSM to summarize how the expected counterfactual happiness Y_a varies as a function of burpees completed a:

$$E_{U,X}(Y_a) = m(a|\beta) = \beta_0 + \beta_1 a$$

with $a \in \{1, 2, 3, 4, 5, 6, 7\}$. For simplicity, we are treating this MSM as the truth. This specification assumes a linear change in the expected counterfactual happiness with increasing burpees completed. (Alternatively, we can consider this a working MSM and the parameter as the projection of the true causal curve onto this linear model.)

IPTW for the MSM parameter without stabilized weights:

1. Estimate the treatment mechanism $P_0(A|W) = g_0(A|W)$, which is the conditional probability of completing A burpees, given the rider's characteristics. Use multinomial logistic regression.
Hint: we already did this! Skip to the next step.

2. Predict each rider's probability of her observed exposure (burpees completed), given his/her covariates $g_n(A_i|W_i)$.
Hint: we already did this! Skip to the next step.

3. Create the vector `wt` as the inverse of the predicted probabilities.
Hint: we already did this! Skip to the next step.

4. Estimate the parameters of the MSM by regressing the observed outcome Y on the exposure A according to the MSM. You must specify the `weights` and the `data`.

5. Interpret the results.

6 Weight stabilization in IPTW for a MSM parameter

For MSMs without effect modification $m(a|\beta)$, a common choice for the numerator of the weights $g^*(A)$ is the marginal probability of the exposure A:

$$st.wt_i = \frac{g_n^*(A)}{g_n(A|W)}, \text{ where } g_n^*(A) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}(A_i = a)$$

For rare exposures, both the numerator and denominator will be small, leading to less extreme weights and more efficient estimators.

The positivity assumption for a parameter defined with a non-saturated MSM depends on the choice of the numerator:

$$sup_{a \in \mathcal{A}} \frac{g^*(a)}{g_0(a|w)} < \infty \text{ for all } w \text{ for which } P_0(W = w) > 0$$

By choosing the numerator to be the marginal probability of each exposure level, then we only need that any value of the exposure, occurring with a non-zero probability, must also occur with a non-zero probability within all possible strata of baseline covariates. Consequently, the statistical estimand is still defined if some levels of the exposure occur with zero probability. For instance, in this study there were no riders completing 3.5 burpees. By using an MSM, we can smooth over exposure levels.

Note: If we consider the MSM to be a "working" model, then the choice of the numerator changes the target parameter. (See lecture notes for more details)

Implement IPTW for a MSM parameter with stabilized weights

1. Estimate the treatment mechanism $g_0(A|W) = P_0(A|W)$. *Hint: We already did this! Skip to next step.*
2. Predict the probability of the observed exposure for each ride `gAW`. *Hint: We already did this! Skip to next step.*
3. Create the stabilized weights `wt.MSM`:

$$st.wt_i = \frac{g_n^*(A)}{g_n(A|W)}, \text{ where } g_n^*(A) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}(A_i = a)$$

   (a) Create empty vector `gA` of length n for the numerator of the weights.

   (b) Index the vector `gA` by exposure status and assign the appropriate marginal probability.
   Hint: For riders completing $A = 1$ burpee, the numerator $g^*(A)$ is the observed proportion with $A = 1$.

```
> gA[ObsData$A==1] <- mean(ObsData$A==1)
> # Implement the analogous code for exposure levels A=2,..., A=7
```

   (c) Create the stabilized weight:

```
> wt.MSM <- gA/gAW
```

   (d) Look at the distribution of the stabilized weights.

4. Estimate the parameters of the MSM by regressing the observed outcome Y and on the exposure A. You must specify the `weights` and the `data`.
5. Are the estimated parameters of the MSM the same?