R Assignment 4- Part B

PH252D Fall 2013
Introduction to Causal Inference

Assigned: November 20, 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

2 Implement TMLE for the G-Computation estimand

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 ($A = 1$):

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

Under the necessary causal assumptions (i.e. the backdoor criteria and the positivity assumption), the corresponding statistical estimand is given by the G-Computation formula:

$$\Psi(P_0) = E_0\big[\bar{Q}_0(7, W) - \bar{Q}_0(1, W)\big]$$

1. Set the seed to 252.
2. Import the data set `Rassign4.Fa2013.csv` and assign it to object `FullData`.
3. Create data frame `ObsData`, consisting riders completing $A = 1$ or $A = 7$ burpees:

```
> ObsData<- FullData[FullData$A==1 | FullData$A==0, ]
```

4. Assign the number of riders in `ObsData` to `n`.
5. Create a new exposure variable `A.binary`, which equals 1 for riders completing $A = 7$ burpees and equals 0 for riders completing $A = 1$ burpee.
6. Use the `table` function to make sure your code is correct.
7. Implement `tmle` using SuperLearner with the default library for initial estimation of $\bar{Q}_0(A, W)$ and $g_0(A|W)$. Be sure to specify the outcome (`Y=ObsData$Y`), the exposure (`A=A.binary`) and the covariates (`W=subset(ObsData, select=c(W1,W2))`).
8. Comment on the point estimates from IPTW (part A) and from TMLE.

3 Evaluate the finite sample performance of TMLE

1. Set the seed to `set.seed(252)`. Create a vector `estimates` of size `R=500`.
2. Within a `for` loop, repeat the following `R=500` times.
   (a) Draw a sample of size 5000 independently from data generating process given in Section 2 of R assignment 4.
   (b) Create data frame `ObsData`, consisting riders completing $A = 1$ or $A = 7$ burpees.
   (c) Assign the number of riders in `ObsData` to `n`.
   (d) Create a new exposure variable `A.binary`, which equals 1 for riders completing $A = 7$ burpees and equals 0 for riders completing $A = 1$ burpee.
   (e) Implement `tmle` using SuperLearner with the default library for initial estimation of $\bar{Q}_0(A, W)$ and $g_0(A|W)$.
   (f) Save the resulting point estimate as a row in `estimates`
3. What is the average estimate? What is the bias, average deviation from the point estimate and the true value? How variable are the estimates?
4. Create a histogram of the point estimates.
   Hint: Use `hist` function.