R Assignment 4 - IPTW

Introduction to Causal Inference

Assigned: April 4, 2018
Write-ups due: April 16, 2018 in class. 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 and Causal Roadmap

"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

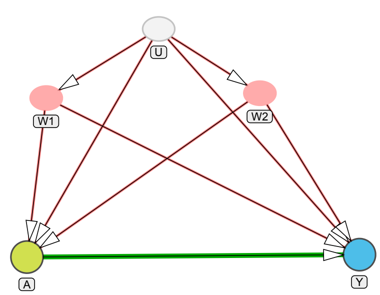

Figure 1: DAG corresponding to the study of burpees and happiness among BART riders.

Causal Roadmap Rundown

This is a very, very quick summary for review. Each step of the roadmap 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 $\mathcal{M}^{\mathcal{F}}$:**
   - 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: $U = (U_{W1}, U_{W2}, U_A, U_Y) \sim \mathbb{P}_U$. 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:

$$\Psi^{\mathcal{F}}(\mathbb{P}_{U,X}) = \mathbb{E}_{U,X}(Y_a) - \mathbb{E}_{U,X}(Y_{a'})$$

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

   Alternatively, we could also consider a marginal structural model (MSM) to summarize how the counterfactual mean happiness $\mathbb{E}_{U,X}(Y_a)$ changes as a function burpees completed a. Consider, for example,

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

   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:

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

   where $g^*(a)$ specifies how much weight we put on specific values of a and where $m(a|\beta) = \beta_0 + \beta_1 a$. For an MSM without effect modifiers (as above), the typical choice for $g^*(a)$ is the marginal probability of the exposure $\mathbb{P}_0(A = a)$.

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 $\mathcal{M}^{\mathcal{F}}$. This yields n i.i.d. copies of random variable $O = (W1, W2, A, Y) \sim \mathbb{P}_0$. The statistical model $\mathcal{M}$ for the set of allowed observed data distributions is non-parametric.

5. **Assess identifiability:**
   In the original SCM $\mathcal{M}^{\mathcal{F}}$, 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 $\mathcal{M}^{\mathcal{F}^*}$ to denote the original SCM augmented by the assumptions needed for identifiability. Under $\mathcal{M}^{\mathcal{F}^*}$, the backdoor criteria will hold conditionally on the set of baseline covariates $W = (W1, W2)$.

   To identify $\mathbb{E}_{U,X}(Y_a)$ (the expected counterfactual happiness under a given burpee level) with the G-Computation formula, we also need the positivity assumption to hold

   $$min_{a \in \mathcal{A}}\ \mathbb{P}_0(A = a|W1 = w1, W2 = w2) > 0$$

   for all $(w1, w2)$ for which $\mathbb{P}_0(W1 = w1, W2 = w2) > 0$. 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 weight function $g^*(A)$.

6. **Specify the statistical estimand:**
   Under the working SCM $\mathcal{M}^{\mathcal{F}^*}$ and with the positivity assumption, the average treatment effect $\Psi^{\mathcal{F}}(\mathbb{P}_{U,X}) = \mathbb{E}_{U,X}(Y_a - Y'_a)$ can be identified with the G-Computation formula:

   $$\Psi(\mathbb{P}_0) = \mathbb{E}_{0,W}\left[\mathbb{E}_0(Y|A = a, W) - \mathbb{E}_0(Y|A = a', W)\right]$$

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

   For the target causal parameter defined with a working MSM, the statistical parameter is the solution in β to the projection of the conditional mean outcome of the observed outcome $\mathbb{E}_0(Y|A = a, W)$ onto the MSM $m(a|\beta)$.

2 Import and explore data set `RAssign4.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) **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.*
   (b) **Comment.**

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

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 ($A = 1$):

$$\Psi^{\mathcal{F}}(\mathbb{P}_{U,X}) = \mathbb{E}_{U,X}(Y_7) - \mathbb{E}_{U,X}(Y_1)$$

1. **We need to estimate the treatment mechanism $\mathbb{P}_0(A|W) = g_0(A|W)$, 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 $\hat{g}(A_i|W_i)$:**

   (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. Comment on the distribution of weights.**

4. **Evaluate the IPTW estimand:**

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

   The first quantity is the weighted mean outcome, where riders completing $A_i = 7$ burpees receive weight $1/\hat{g}(A_i = 7|W_i)$ 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 $1/\hat{g}(A_i = 1|W_i)$ 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):**

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

6. **Interpret the point estimates.**

4 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 ($A = 1$) on happiness. Now suppose we want to smooth over levels of the exposure. Consider the following MSM to summarize how the expected counterfactual happiness $\mathbb{E}_{U,X}(Y_a)$ varies as a function of burpees completed a:

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

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, where all levels of the exposure are weighted evenly.)

4.1 IPTW for the MSM parameter without stabilized weights:

1. **Estimate the treatment mechanism $\mathbb{P}_0(A|W) = g_0(A|W)$, 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 $\hat{g}(A_i|W_i)$.**
   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 corresponding to the MSM by regressing the observed outcome Y on the exposure A according to $m(a|\beta)$. You must specify the `weights` and the `data`.**
5. **Interpret the results.**

4.2 IPTW for a MSM parameter with Stabilized Weights

For MSMs without effect modification (e.g. $m(a|\beta) = \beta_0 + \beta_1 a$), a common choice for the numerator of the weights $g^*(A)$ is the marginal probability of the exposure $\mathbb{P}_0(A = a)$. Therefore, our stabilized weights are

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

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:

$$sup_{a\in\mathcal{A}} \frac{g^*(a)}{g_0(a|w)} < \infty \text{ for all } w \text{ for which } \mathbb{P}_0(W = w) > 0$$

By choosing the numerator as the marginal probability of each exposure level, 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 the Lecture Notes and the Appendix for more details)

4.3 Implement IPTW for a MSM parameter with stabilized weights

1. **Estimate the treatment mechanism $g_0(A|W) = \mathbb{P}_0(A|W)$.** *Hint: We already did this! Skip to next step.*
2. **Predict the probability of the observed exposure for each rider `gAW`.** *Hint: We already did this! Skip to next step.*
3. **Create the stabilized weights `wt.MSM`:**

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

   (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 estimated 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) **Comment on the distribution of the stabilized weights.**

4. **Estimate the parameters corresponding to 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 coefficients the same? Briefly discuss.**

5 Improving the IPTW Estimator the G-computation formula

This section uses the data generating distribution given in `Rassign4_modifiedIPTW.R`. In particular, the data generating distribution $\mathbb{P}_0$ is given by

$$W \sim Bernoulli(1/2)$$
$$A|W \sim Bernoulli(0.2 + 0.6 \times W)$$
$$Y|A, W =_d 1000 + \mathbb{I}(\tilde{U} < logit^{-1}(W \times A)),$$

where $\tilde{U} \sim Uniform(0, 1)$ is independent of the other variables and where $=_d$ indicates "has the same distribution as". Note that Y only takes on the values 1000 and 1001.

Our goal is to estimate $\Psi(\mathbb{P}_0) = \sum_w \mathbb{E}_0[Y|A = 1, W = w]\mathbb{P}_0(W = w)$. The file `Rassign4_modifiedIPTW.R` also contains a wrapper function which implements the IPTW estimator and modified Horvitz-Thompson estimator (i.e. stablized IPTW) of $\Psi(\mathbb{P}_0)$. In this problem we will assume that g_0 is known to the investigators (as in a randomized controlled trial without missingness). These estimators are then given by:

$$\hat{\Psi}_{IPTW}(\mathbb{P}_n) = \frac{1}{n}\sum_{i=1}^{n} \frac{A_i}{g_0(1|W_i)} Y_i$$
$$\hat{\Psi}_{HT}(\mathbb{P}_n) = \frac{\sum_{i=1}^{n} \frac{A_i}{g_0(1|W_i)} Y_i}{\sum_{i=1}^{n} \frac{A_i}{g_0(1|W_i)}}.$$

In class we discussed how the modified Horvitz-Thompson estimator will often (although not always) yield finite sample improvements to the standard IPTW estimator. We also alluded to the fact that the modified Horvitz-Thompson is asymptotically the same as the standard IPTW estimator; so the two will have similar behavior in large samples.

The code in `Rassign4_modifiedIPTW.R` also contains a space for `my.est`, an estimator that you will define and implement in this section. In particular, we will seek to modify the standard IPTW estimator in a different way to yield both finite sample and asymptotic improvements. The goal is for you to come up with (at least a precursor to) this estimator on your own. In the solution key, we will present the best possible modification to the IPTW estimator in terms of asymptotic performance. In class we will see that this estimator is asymptotically equivalent to the targeted maximum likelihood estimator (TMLE) for $\Psi(\mathbb{P}_0)$. Nonetheless, we expect the TMLE to perform better in finite samples for reasons that will be described in class.

Please complete Questions 1 through 7 listed below.

1. **Run the code given in `Rassign4_modifiedIPTW.R` and report how the standard IPTW and modified Horvitz-Thompson estimators perform in terms of bias, variance, and MSE. Which**

estimator would you use in practice?
`Note 1:` The estimator `my.est` will return `NA`, because you haven't implemented it yet!
`Note 2:` Both of these estimators are unbiased in finite samples when g_0 is known; so any estimated bias is the result of only taking a finite number of Monte Carlo draws. To test this, try increasing the number of Monte Carlo draws, or simply repeating the Monte Carlo simulation of size 2000 multiple times.

2. **Look at the IPTW column in the `est` matrix. What do you notice about the IPTW estimates across these 2000 Monte Carlo draws?**
   `Hint:` Recall the values that Y can take.

One way to address the problem that you described above is to use a modified Horvitz-Thompson estimator, which automatically respects the bounds of the model and can lead to better finite sample performance. Nonetheless, there is another valid way to address this problem. Note that the IPTW estimate is an average of terms which are 0 (people with $A = 0$) and of terms which are larger than 1000 (people with $A = 1$). The calculations in the next set of questions will be useful for developing the intuition needed to create your own estimator.

3. **What is the variance of a random variable X with $Pr(X = 0) = 1/2$ and $Pr(X = 1) = 1/2$?**

4. **What is the variance of a random variable $X2$ with $Pr(X2 = 0) = 1/2$ and $Pr(X2 = 1000) = 1/2$?**

5. **Can you think of how the above two calculations could be relevant to improving the IPTW estimator in this problem?**
   `Hint:` We currently have an estimator that is an empirical mean of variables like those in Question 4. Is there some transformation of the outcome Y that you can do to make your estimator behave more like an empirical mean of variables like those in Question 3?

6. ***Bonus:*** **Write down an estimator $\hat{\Psi}_{my.est}$ which applies the ideas of the previous three questions into an estimator. There's no need to give the best possible estimator, but you should give an estimator that outperforms the IPTW estimator by a significant margin (i.e. does as or almost as well as the modified Horvitz-Thompson estimator in terms of bias/variance/MSE).**

7. ***Bonus:*** **Code your estimator and replace the NA on the line `my.est = NA` with the estimator you defined in the previous question. Report the bias/variance/MSE of your estimator over the 2000 Monte Carlo draws.**