R Assignment 2

Introduction to Causal Inference

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. You need to bring your own completed assignment to class.

1 "Time to prevent child malnutrition in Sahel"

Excerpted from http://www.irinnews.org/report/98941/time-to-prevent-child-malnutrition-in-sahel

"DAKAR, 14 October 2013 (IRIN) - Malnutrition among children under age five in the Sahel is expected to rise again this year, despite decent rains and more or less average harvest predictions... There are multiple reasons malnutrition cases have risen this year, including high food prices, conflict, high incidence rates of malaria and improved humanitarian coverage - which may mean better reporting of child malnutrition. Other structural causes include weak health systems, deep poverty, poor water and sanitation conditions, and inadequate infant care practices, according to the health and nutrition NGO Alima...

At Konseguela health post, in Koutiala District in Mali's Sikasso Region, MSF set out to prevent malnutrition by addressing the gamut of related causes. As part of a two-year programme, the organization gave all children antimalarial tablets - whether or not they had the disease - during the four-month malaria season. They also handed out mosquito nets, made rapid malaria tests available and taught community workers how to measure weight loss using arm-circumference measures... The programme also vaccinated children against pneumococcal diseases, administered oral rehydration salts to children with diarrhoea, dispensed chlorine for water treatment, and offered nutritional supplements and regular free follow-up visits from a health worker. Since the programme started two years ago, stunting in Konseguela has fallen by one-third and child mortality by half."

- Suppose we are interesting in evaluating the effect of this integrated approach on all-cause childhood mortality in the greater Sahel region. Let $W1$ be an indicator of the child's access to health care facilities. Let $W2$ be an indicator, equaling 1 if the child lives in area with recent conflict. The intervention A is also an indicator, equaling 1 if the child received prevention package and equaling 0 if the child received the standard of care. Finally, the outcome Y represents the child's survival status at the end of two years.
- This *simplified* study can be translated into the following structural causal model (SCM) $\mathcal{M}^F$:
  - Endogenous nodes: $X = (W1, W2, A, Y)$
  - Background (exogenous) variables: $U = (U_{W1}, U_{W2}, U_A, U_Y) \sim P_U$
  - Structural equations F:

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

- The target causal parameter is the difference in the counterfactual probability of survival if all children received the combination prevention package and the counterfactual probability of survival if all children did not receive the package:

$$\Psi^F(P_{U,X}) = P_{U,X}(Y_1 = 1) - P_{U,X}(Y_0 = 1) = E_{U,X}(Y_1) - E_{U,X}(Y_0)$$

- We assume the observed data $O = (W1, W2, A, Y) \sim P_0$ were generated by sampling n independent times from a data generating system contained in $\mathcal{M}^F$.
- The target causal quantity is not identified under the backdoor criterion in the SCM $\mathcal{M}^F$. A sufficient, but not minimal, identifiability assumption is that all exogenous errors 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) $U_Y \perp\!\!\!\perp U_{W1},\ U_Y \perp\!\!\!\perp U_{W2}$.
- Under the working SCM $\mathcal{M}^{F*}$, the average treatment effect $\Psi^F(P_{U,X})$ is identified using the G-Computation formula:

$$\begin{aligned}\Psi(P_0) &= E_0\big[E_0(Y|A=1,W1,W2) - E_0(Y|A=0,W1,W2)\big] \\ &= \sum_{w1,w2}\big[E_0(Y|A=1,W1=w1,W2=w2) - E_0(Y|A=0,W1=w1,W2=w2)\big]P_0(W1=w1,W2=w2)\end{aligned}$$

The statistical estimand $\Psi(P_0)$ is the difference in the strata-specific conditional probability of survival under the intervention and control, averaged with respect to the distribution of baseline covariates (health care access and conflict history).

For the statistical estimand to be well-defined, we need an additional statistical assumption on variability of the exposure within covariate strata. This condition is known as the positivity assumption:

$$\min_{a\in\mathcal{A}} g_0(A=a|W_1=w_1,W_2=w_2) > 0, \text{ for all } w \text{ where } P_0(W_1=w_1,W_2=w_2) > 0$$

where $\mathcal{A}$ denotes the set of exposures of interest and $g_0(A|W_1,W_2) = P_0(A|W_1,W_2)$ denotes the conditional distribution of the exposure, given the covariates.

- Letting $W = (W1, W2)$ represent the covariates in the adjustment set, we can re-express this identifiability result as

$$\Psi(Q_0) = E_0\big[\bar{Q}_0(1,W) - \bar{Q}_0(0,W)\big]$$

where $Q_0 = (\bar{Q}_0(A,W), Q_{0,W})$ is the relevant part of the likelihood and consists of the conditional mean outcome given the intervention and baseline covariates $E_0(Y|A,W)$ and the marginal distribution of baseline covariates $P_0(W)$.

2 A specific data generating process

Consider a specific data generating process (unknown to the researchers), which is one of many compatible with the SCM $\mathcal{M}^F$. The endogenous factors U are independently generated as

$$\begin{aligned}U_{W1} &\sim Uniform(0,1)\\ U_{W2} &\sim Uniform(0,1)\\ U_A &\sim Uniform(0,1)\\ U_Y &\sim Uniform(0,1)\end{aligned}$$

Given the exogenous U, the endogenous variables are deterministically generated as

$$\begin{aligned}W1 &= \mathbb{I}[U_{W1} < 0.50]\\ W2 &= \mathbb{I}[U_{W2} < 0.50]\\ A &= \mathbb{I}[U_A < expit(-0.5 + W1 - 1.5^*W2)]\\ Y &= \mathbb{I}[U_Y < expit(-0.75 + W1 - 2^*W2 + 2.5^*A + A^*W1)]\end{aligned}$$

1. **Evaluate the postivity assumption in closed form for this data generating process.**
   Hints: For the positivity assumption to hold, there must be a positive probability of receiving the

intervention package ($A = 1$) and the standard of care ($A = 0$) within all possible strata of health care access ($W1$) and conflict history ($W2$). Since the treatment is binary, we need to check if

$$0 < P_0(A = 1|W1 = 1, W2 = 1) < 1$$
$$0 < P_0(A = 1|W1 = 1, W2 = 0) < 1$$
$$0 < P_0(A = 1|W1 = 0, W2 = 1) < 1$$
$$0 < P_0(A = 1|W1 = 0, W2 = 0) < 1$$

This would imply that the probability of receiving the standard of care ($A = 0$) within all strata of baseline covariates would also be bounded between 0 and 1.

In this particular data generating system (one of many compatible with the SCM), the conditional probability of receiving the intervention given the baseline covariates is

$$P_0(A = 1|W1, W2) = expit(-0.5 + W1 - 1.5^*W2)$$

2. ***Bonus (Optional):* Evaluate the statistical estimand $\Psi(P_0)$ in closed form for this data generating process.**
Hints: In this particular data generating system (one of many compatible with the SCM), the conditional probability of survival, given the intervention and the baseline covariates is

$$P_0(Y = 1|A, W1, W2) = E_0(Y|A, W1, W2) = expit(-0.75 + W1 - 2^*W2 + 2.5^*A + A^*W1)$$

The marginal distribution of $W1$ (access to healthcare) is Bernoulli with probability 0.5:

$$P_0(W1 = 1) = E_0(W1) = 0.5$$

The marginal distribution of $W2$ (recent conflict near the child's home) is Bernoulli with probability 0.5:

$$P_0(W2 = 1) = E_0(W2) = 0.5$$

Since the two baseline covariates are *independent*, their joint distribution is

$$P_0(W1 = 1, W2 = 1) = P_0(W1 = 1)^*P_0(W2 = 1) = 0.5^*0.5 = 0.25$$

3 Translate this data generating process into simulations.

1. **First set the seed to 252.**
2. **Set the number of draws $n = 100,000$.**
3. **Sample n i.i.d. observations of random variable $O = (W1, W2, A, Y) \sim P_0$.** Recall the *expit* function is given by the `plogis` function in `R`.
4. ***Bonus*: Intervene to set the exposure to the combination package ($A = 1$) and generate the counterfactual outcome Y_1. Intervene to set the exposure to the standard of care ($A = 0$) and generate the counterfactual outcomes Y_0. Evaluate the causal parameter $\Psi^F(P_{U,X})$.**
5. **Evaluate the positivity assumption.** Take the mean of the exposure A in each strata of covariates $(W1, W2)$.
6. **Evaluate the statistical estimand $\Psi(P_0)$ and assign the value ψ_0 to `Psi.P0`.**
7. **Interpret $\Psi(P_0)$.**

4 The simple substitution estimator based on the G-Computation formula

We usually do not know the true distribution of the observed data P_0. Instead, we only have a sample of n i.i.d. observations of O. The empirical distribution, denoted P_n, puts weight $1/n$ on each observation O_i. An intuitive estimator of the statistical estimand is the simple substitution estimator based on the G-Computation formula. Briefly, the algorithm estimates the relevant parts of the observed data distribution, denoted Q_0, and plugs them into the parameter mapping Ψ:

$$\hat{\Psi}(P_n) = \frac{1}{n}\sum_{i=1}^{n}\left[\bar{Q}_n(1, W_i) - \bar{Q}_n(0, W_i)\right]$$

where $\bar{Q}_n(A, W)$ denotes an estimate of the conditional mean function $\bar{Q}_0(A, W) = E_0(Y|A, W)$, and the sample proportion has been used to estimate marginal distribution of covariates $Q_{0,W} = P_0(W)$.

As in `R` lab 2, we will use simulations to evaluate the performance of the simple substitution estimator, when various parametric regression models are assumed to estimate $\bar{Q}_0(A, W)$. Specifically, for $R = 500$ iterations, we will sample $n = 200$ i.i.d. observations from P_0, implement 4 estimators and save the resulting point estimates.

1. **Set the number of iterations `R` to 500 and the number of observations `n` to 200. Do *not* reset the seed.**

2. **Create a $R = 500$ by 4 matrix `estimates` to hold the resulting estimates obtained at each iteration.** The rows will correspond to iterations and the columns to different estimators of $\bar{Q}_0(A, W)$.

```
> # Hint: the following code creates an matrix filled with NA of size 10 by 10
> estimates<- matrix(NA, nrow=10, ncol=10)
```

3. **Inside a `for` loop from `r` equals 1 to `R` (500), do the following.**

   (a) Sample n i.i.d. observations of $O = (W1, W2, A, Y)$.

   (b) Create a data frame `Obs` of the resulting observed data.

   (c) Copy the data set `Obs` into two new data frames `txt` and `control`. Then set `A=1` for all units in `txt` and set `A=0` for all units in the control.

   (d) Estimator #1: Use `glm` function to estimate $\bar{Q}_0(A, W)$ (the conditional probability of survival, given the intervention and baseline covariates) based on the following parametric regression model:

$$\bar{Q}_0^1(A, W) = expit(\beta_0 + \beta_1 A)$$

   Be sure to specify the arguments `family='binomial'` and `data=Obs`.

   (e) Estimator #2: Use `glm` function to estimate $\bar{Q}_0(A, W)$ based on the following parametric regression model:

$$\bar{Q}_0^2(A, W) = expit(\beta_0 + \beta_1 A + \beta_2 W1)$$

   Be sure to specify the arguments `family='binomial'` and `data=Obs`.

   (f) Estimator #3: Use `glm` function to estimate $\bar{Q}_0(A, W)$ based on the following parametric regression model:

$$\bar{Q}_0^3(A, W) = expit(\beta_0 + \beta_1 A + \beta_2 W2)$$

   Be sure to specify the arguments `family='binomial'` and `data=Obs`.

   (g) Estimator #4: Use `glm` function to estimate $\bar{Q}_0(A, W)$ based on the following parametric regression model:

$$\bar{Q}_0^4(A, W) = expit(\beta_0 + \beta_1 A + \beta_2 W1 + \beta_3 W2 + \beta_4 A^* W1 + \beta_5 A^* W2)$$

   Be sure to specify the arguments `family='binomial'` and `data=Obs`.

(h) For *each* estimator of $\bar{Q}_0(A, W)$, use the `predict` function to get the expected (mean) outcome for each unit under the intervention $\bar{Q}_n(1, W_i)$. Be sure to specify the arguments `newdata=txt` and the `type='response'`.

(i) For *each* estimator of $\bar{Q}_0(A, W)$, use the `predict` function to get the expected (mean) outcome for each unit under the control $\bar{Q}_n(0, W_i)$. Be sure to specify the arguments `newdata=control` and the `type='response'`.

(j) For *each* estimator of $\bar{Q}_0(A, W)$, estimate of $\Psi(P_0)$ by substituting the predicted mean outcomes under the treatment $\bar{Q}_n(1, W_i)$ and control $\bar{Q}_n(0, W_i)$ into the G-Computation formula and using the sample proportion to estimate the marginal distribution of baseline covariates:

$$\hat{\Psi}(P_n) = \frac{1}{n} \sum_{i=1}^{n} \left[\bar{Q}_n(1, W_i) - \bar{Q}_n(0, W_i) \right]$$

(k) Assign the resulting values as a row in matrix `estimates`.

```
> # Hint:  the following code assigns the 4 resulting estimates
> # (denoted psi.hat1, psi.hat2, psi.hat3, psi.hat4) from iteration r to row r
> estimates[r,] <- c(psi.hat1, psi.hat2, psi.hat3, psi.hat4)
```

Some additional hints:

- See `R` lab 2 for an example implementation of the simple substitution estimator and a `for` loop. Here, we are evaluating 4 estimators simultaneously.
- While you are writing your code and testing it, set the number of iterations `R` to a smaller number (e.g. 5). This will help save time.
- If you get stuck, talk to your classmates and GSIs.

5 Performance of the estimators.

1. **What is the average value of each estimator of $\Psi(P_0)$ across $R = 500$ simulations?**
   Hint: Take the mean of each column of `estimates`.

2. **Estimate the bias of each estimator.** Hint: For each estimator, average the difference between point estimate ψ_n and the truth ψ_0.

$$Bias\big(\hat{\Psi}(P_n)\big) = E_0\big[\hat{\Psi}(P_n) - \Psi(P_0)\big]$$

3. **Estimate the variance of each estimator.**

$$Variance\big(\hat{\Psi}(P_n)\big) = E_0\left(\left(\hat{\Psi}(P_n) - E_0[\hat{\Psi}(P_n)]\right)^2\right)$$

   Hint: use the `var` function.

4. **Estimate the mean squared error of each estimator.** On average, how far is the estimator from the truth?

$$MSE\big(\hat{\Psi}(P_n)\big) = E_0\left(\left(\hat{\Psi}(P_n) - \Psi(P_0)\right)^2\right) = Bias^2 + Var$$

5. **Briefly comment on the performance of the estimators. Which estimator has the lowest MSE over the $R = 500$ iterations? Are you surprised?**