R Assignment 2

Laura B. Balzer

Biostat683 - Intro. to Causal Inference

Assigned: October 20, 2021
Write-ups due: Uploaded to your personal GoogleDrive folder by November 1, 2021 by 2:30pm. 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. Use of RMarkdown is strongly encouraged.

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."

- **Step 0: Scientific question:** 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 that the child lives conflict area. Let $W2$ be an indicator that the child has access to health care. The intervention A is also an indicator variable, equaling 1 if the child subsequently received prevention package and equaling 0 if the child received the standard-of-care. Finally, the outcome Y is an indicator that the child survived through the two years of follow-up.
- **Step 1: Causal model representing real knowledge:** Suppose this *simplified* study can be translated into the following structural causal model (SCM) $\mathcal{M}^*$:
  - Endogenous nodes: $X = (W1, W2, A, Y)$
  - Background variables: $U = (U_{W1}, U_{W2}, U_A, U_Y) \sim \mathbb{P}_U$
  - Structural equations F:

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

- **Step 2: Counterfactuals & causal parameter:** 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:

$$\Psi^*(\mathbb{P}^*) = \mathbb{E}^*(Y_1) - \mathbb{E}^*(Y_0) = \mathbb{P}^*(Y_1 = 1) - \mathbb{P}^*(Y_0 = 1)$$

2 Roadmap Questions

1. **Step 3: Observed data & link to causal:** Suppose the observed data consist of n independent, identically distributed (i.i.d) draws of the random variable $O = (W1, W2, A, Y)$.
   (a) Specify the link between the SCM and the observed data.
   (b) What restrictions, if any, does the SCM place on the set of allowed distributions for the observed data?
   (c) What notation do we use to denote the true (but unknown) distribution of the observed data and the statistical model?
2. **Step 4-5: Identification & statistical estimand:**
   (a) Using the backdoor criterion, assess identifiability.
   (b) If the target causal parameter is not identified, under what assumptions would it be?
   (c) Specify the target parameter of the observed data distribution (i.e., the statistical estimand). Interpret it.
   (d) What is the relevant positivity assumption? Is it reasonable here?

Solution:

1. Step 3
   (a) We assume the observed data $O = (W1, W2, A, Y)$ were generated by sampling n i.i.d. times from a data generating system compatible with $\mathcal{M}^*$. This provides a link between the causal model $\mathcal{M}^*$ and the observed data O. The distribution of the background variables U and the structural equations F identify the distribution of the endogenous variables X and thus the distribution of the observed data O, which is denoted $\mathbb{P}_0$.
   (b) We have not placed any restrictions on the statistical model $\mathcal{M}$, which is thereby non-parameteric.
   (c) We use $\mathbb{P}_0$ to denote the true (but unknown) distribution of the observed data, and $\mathcal{M}$ to denote the statistical model.
2. Steps 4-5:
   (a) In the SCM $\mathcal{M}^*$, the target causal quantity is not identified under the backdoor criterion. In other words, we cannot express $\Psi^*(\mathbb{P}^*)$ as a parameter of the observed data distribution $\mathbb{P}_0$ without additional (unsubstantiated) assumptions.
   (b) A sufficient, but not minimal, identifiability assumption is that all background 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) $U_Y \perp\!\!\!\perp U_{W1}$, $U_Y \perp\!\!\!\perp U_{W2}$.
   (c) Despite lack of identifiability, we can still "commit" to an interesting statistical estimand inspired by our scientific/causal question:

$$\begin{aligned}\Psi(\mathbb{P}_0) &= \mathbb{E}_0\big[\mathbb{E}_0(Y|A=1,W1,W2) - \mathbb{E}_0(Y|A=0,W1,W2)\big] \\ &= \sum_{w1,w2} \big[\mathbb{E}_0(Y|A=1,w1,w2) - \mathbb{E}_0(Y|A=0,w1,w2)\big]\mathbb{P}_0(w1,w2)\end{aligned}$$

The statistical estimand $\Psi(\mathbb{P}_0)$ is the difference in the strata-specific conditional probability of survival under the intervention and standard-of-care, averaged with respect to the distribution of baseline covariates (health care access and conflict history).

Formally, the parameter Ψ is a mapping from the statistical model $\mathcal{M}$ to the parameter space $\mathbb{R}$. In other words, it is a function that takes as input any distribution compatible with the statistical model $\mathcal{M}$ and gives as output a value in the paramter space.

(d) For the statistical estimand to be well-defined, we need variability of the exposure (receiving the prevention package or not) within adjustment strata (healthcare access and conflict history).

This condition is known as the positivity assumption:

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

where $\mathcal{A}$ denotes the set of exposures of interest $\{0, 1\}$.

Theoretically, the positivity assumption is probably reasonable here. Within the target catchment of the greater Sahel region, there is likely to be a positive probability of receiving the prevention package or not within all strata of healthcare access and conflict history. Practically, there might be violations if, for example, conflict prohibits delivery of the prevention package or if areas with strong healthcare at baseline are deprioritized for the progam.

3 A specific data generating process

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

$$\begin{aligned} W1 &= \mathbb{I}[U_{W1} < 0.50] \\ W2 &= \mathbb{I}[U_{W2} < 0.50] \\ A &= \mathbb{I}[U_A < logit^{-1}(-0.5 + W1 - 1.5^* W2)] \\ Y &= \mathbb{I}[U_Y < logit^{-1}(-0.75 + W1 - 2^* W2 + 2.5^* A + A^* W1)] \end{aligned}$$

1. **Evaluate the postivity assumption in closed form for this data generating process.**

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

   $$\mathbb{P}_0(A = 1 | W1, W2) = logit^{-1}(-0.5 + W1 - 1.5^* W2)$$

2. ***Bonus (Optional):* Evaluate the statistical estimand $\Psi(\mathbb{P}_0)$ in closed form for this data generating process.**

Solution:

1. In this example, we know the conditional distribution of the intervention given the baseline covariates:

$$\mathbb{P}_0(A = 1|W1, W2) = logit^{-1}[-0.5 + W1 - 1.5^*W2]$$

Therefore, we can plug in different values for $W1$ and $W2$ and can calculate the relevant probabilities directly:

$$\mathbb{P}_0(A = 1|W1 = 1, W2 = 1) = logit^{-1}[-0.5 + 1 - 1.5^*1] = 0.269$$
$$\mathbb{P}_0(A = 1|W1 = 1, W2 = 0) = logit^{-1}[-0.5 + 1 - 1.5^*0] = 0.622$$
$$\mathbb{P}_0(A = 1|W1 = 0, W2 = 1) = logit^{-1}[-0.5 + 0 - 1.5^*1] = 0.119$$
$$\mathbb{P}_0(A = 1|W1 = 0, W2 = 0) = logit^{-1}[-0.5 + 0 - 1.5^*0] = 0.378$$

Since the relevant probabilities are between 0 and 1, there are no theoretical violations of the positivity assumption.

```
> # relevant R code
> probA.givenW <- function(W1, W2){
+   plogis(-.5 + W1 -1.5*W2)
+ }
> probA.givenW(1,1)

[1] 0.2689414

> probA.givenW(1,0)

[1] 0.6224593

> probA.givenW(0,1)

[1] 0.1192029

> probA.givenW(0,0)

[1] 0.3775407
```

2. *Bonus:* 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

$$\mathbb{P}_0(Y = 1|A, W1, W2) = \mathbb{E}_0(Y|A, W1, W2) = logit^{-1}(-0.75 + W1 - 2^*W2 + 2.5^*A + A^*W1)$$

The marginal distribution of $W1$ (recent conflict near the child's home) is Bernoulli with probability 0.5:

$$\mathbb{P}_0(W1 = 1) = \mathbb{E}_0(W1) = 0.5$$

The marginal distribution of $W2$ (access to healthcare) is Bernoulli with probability 0.5:

$$\mathbb{P}_0(W2 = 1) = \mathbb{E}_0(W2) = 0.5$$

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

$$\mathbb{P}_0(W1 = 1, W2 = 1) = \mathbb{P}_0(W1 = 1)^*\mathbb{P}_0(W2 = 1) = 0.5^*0.5 = 0.25$$

Plugging into the G-Computation formula, we have

$$\begin{aligned}
\Psi(\mathbb{P}_0) &= \sum_{w1,w2} \big[\mathbb{E}_0(Y|A=1,W1=w1,W2=w2) - \mathbb{E}_0(Y|A=0,W1=w1,W2=w)\big] P(W1=w1,W2=w2) \\
&= \big[\mathbb{E}_0(Y|A=1,W1=1,W2=1) - \mathbb{E}_0(Y|A=0,W1=1,W2=1)\big]\mathbb{P}_0(W1=1,W2=1) \\
&+ \big[\mathbb{E}_0(Y|A=1,W1=1,W2=0) - \mathbb{E}_0(Y|A=0,W1=1,W2=0)\big]\mathbb{P}_0(W1=1,W2=0) \\
&+ \big[\mathbb{E}_0(Y|A=1,W1=0,W2=1) - \mathbb{E}_0(Y|A=0,W1=0,W2=1)\big]\mathbb{P}_0(W1=0,W2=1) \\
&+ \big[\mathbb{E}_0(Y|A=1,W1=0,W2=0) - \mathbb{E}_0(Y|A=0,W1=0,W2=0)\big]\mathbb{P}_0(W1=0,W2=0) \\
&= \big[logit^{-1}(-0.75+1-2^*1+2.5^*1+1^*1) - logit^{-1}(-0.75+1-2^*1+2.5^*0+0^*1)\big](0.5*0.5) \\
&+ \big[logit^{-1}(-0.75+1-2^*0+2.5^*1+1^*1) - logit^{-1}(-0.75+1-2^*0+2.5^*0+0^*1)\big](0.5*0.5) \\
&+ \big[logit^{-1}(-0.75+0-2^*1+2.5^*1+1^*0) - logit^{-1}(-0.75+0-2^*1+2.5^*0+0^*0)\big](0.5*0.5) \\
&+ \big[logit^{-1}(-0.75+0-2^*0+2.5^*1+1^*0) - logit^{-1}(-0.75+0-2^*0+2.5^*0+0^*0)\big](0.5*0.5) \\
&= 0.507
\end{aligned}$$

```
> # in R, E_0(Y| A,W)= plogis(-.75 + W1 - 2*W2 + 2.5*A + A*W1)
> meanY.givenAW <- function(W1, W2, A){
+   plogis(-0.75 + W1 -2*W2 + 2.5*A + A*W1)
+ }
> diff.means <- function(W1, W2){
+   meanY.givenAW(W1,W2, A=1) - meanY.givenAW(W1,W2, A=0)
+ }
> diff.means(1,1)*(0.5*0.5) +
+ diff.means(1,0)*(0.5*0.5) +
+ diff.means(0,1)*(0.5*0.5) +
+ diff.means(0,0)*(0.5*0.5)

[1] 0.506905
```

4 Translate this data generating process into simulations.

1. **First set the seed to 252.**
2. **Write a function to generate the observed data $O = (W1, W2, A, Y)$ and the counterfactual outcomes (Y_1, Y_0).** Recall we generate the counterfactual outcome Y_1 by intervening to set the exposure to the combination package ($A = 1$), and we generate the counterfactual outcomes Y_0 by intervening to set the exposure to the standard-of-care ($A = 0$). Also recall $logit^{-1}$ function is given by the `plogis` function in `R`.
3. **Suppose our target population consists of 100,000 people. Set the number of draws $n = 100,000$. Use your function to generate n i.i.d. observations.**
4. **Does the counterfactual outcome Y_a equal the observed outcome Y when the observed exposure is $A = a$?**
5. ***Bonus*: Evaluate and interpret the causal parameter $\Psi^*(\mathbb{P}^*)$.**

Solution:

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

```
> # 2. Function to generate the observed data and counterfactual outcomes (sometimes call the full data
> generate.data <- function(n){
+   # draw the background factors
+   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 <- runif(n, min=0, max=1)
+   # generate the endogenous variables
+   W1 <- as.numeric(U.W1 < 0.50)
+   W2 <- as.numeric(U.W2 < 0.50)
+   A <- as.numeric(U.A < plogis(-.5 + W1 - 1.5*W2))
+   # structural equation f.Y
+   f.Y <- function(W1,W2,A, U.Y){
+     as.numeric(U.Y < plogis(-.75 + W1 - 2*W2 + 2.5*A + A*W1))
+   }
+   # observed outcome under observed exposure
+   Y <- f.Y(W1=W1, W2=W2, A=A, U.Y=U.Y)
+   # counterfactual outcome under A=1
+   Y1 <- f.Y(W1=W1, W2=W2, A=1, U.Y=U.Y)
+   # counterfactual outcome under A=0
+   Y0 <- f.Y(W1=W1, W2=W2, A=0, U.Y=U.Y)
+
+   data.frame(cbind(W1=W1, W2,A, Y, Y1,Y0))
+
+ }
> # 3. set the number of draws & draw n i.i.d. observations
> n <- 100000
> df <- generate.data(n)

> # 4. consistency assumption
> head(df)

  W1 W2 A Y Y1 Y0
1  0  0 1 1  1  1
2  0  0 1 1  1  0
3  1  0 1 1  1  1
4  0  0 1 1  1  0
5  0  0 1 1  1  1
6  1  0 1 1  1  1

> tail(df)

       W1 W2 A Y Y1 Y0
99995   0  1 0 0  1  0
99996   1  0 1 1  1  1
99997   0  1 0 0  1  0
99998   1  0 1 1  1  1
99999   0  1 0 1  1  1
100000  1  1 0 0  1  0

> txt <- df$A==1; con <- df$A==0
> sum(df[txt, 'Y1'] != df[txt,'Y'])

[1] 0
```

```
> sum(df[con, 'Y0'] != df[con,'Y'])

[1] 0
```

Yes! We have Y_a equals Y when $A = a$

```
> # BONUS:
> # for fun you can generate the counterfactuals and evaluate Psi.F
> Psi.star <-mean(df$Y1) - mean(df$Y0)
> Psi.star

[1] 0.50707
```

The causal risk difference $\Psi^*(\mathbb{P}^*) \approx 0.507$. The difference in the counterfactual probability of survival if all children received the combination intervention package as opposed to the standard-of-care is 50.7%.

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

We usually do not know the true distribution of the observed data $\mathbb{P}_0$, and we do not observe all 100,000 people in our target population. Instead, we only have a finite (small) sample of n i.i.d. observations of O. The empirical distribution, denoted $\mathbb{P}_n$, 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 and plugs them into the parameter mapping Ψ:

$$\hat{\Psi}(\mathbb{P}_n) = \frac{1}{n}\sum_{i=1}^{n}\left[\hat{\mathbb{E}}(Y|A=1,W_i) - \hat{\mathbb{E}}(Y|A=0,W_i)\right]$$

where W denotes the adjustment set; $\hat{\mathbb{E}}(Y|A,W)$ denotes an estimate of the conditional mean outcome $\mathbb{E}_0(Y|A,W)$, and the sample proportion (which simplifies to the empirical mean) has been used to estimate marginal distribution of covariates $\mathbb{P}_0(W)$.

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 $\mathbb{E}_0(Y|A,W)$. For $R = 500$ iterations, we will sample $n = 200$ i.i.d. observations from $\mathbb{P}_0$, implement 4 estimators and save the resulting point estimates. Specifically, we will compare the estimates of $\Psi(\mathbb{P}_0)$ resulting from the following four parametric regressions for estimating $\mathbb{E}_0(Y|A,W)$:

- Regression1: $\mathbb{E}(Y|A,W) = logit^{-1}(\beta_0 + \beta_1 A)$
- Regression2: $\mathbb{E}(Y|A,W) = logit^{-1}(\beta_0 + \beta_1 A + \beta_2 W1)$
- Regression3: $\mathbb{E}(Y|A,W) = logit^{-1}(\beta_0 + \beta_1 A + \beta_2 W2)$
- Regression4: $\mathbb{E}(Y|A,W) = logit^{-1}(\beta_0 + \beta_1 A + \beta_2 W1 + \beta_3 W2 + \beta_4 A^* W1 + \beta_5 A^* W2)$

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 point estimates obtained at each iteration.** The rows will correspond to iterations and the columns to different estimators.

   ```
   > # 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) Use your function from Part 4 to generate n i.i.d. observations. Subset the resulting data.frame to only include the observed data $O = (W1, W2, A, Y)$, and name it `Obs`.

   ```
   > # Hint: if my function from Part4 was called generate.data, then the following would
   > # generate n observataions of (W1,W2,A,Y,Y1,Y2) and subset on the observed data
   > df <- generate.data(n)
   > Obs <- subset(df, select=c(W1,W2,A,Y))
   ```

   (b) 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.

   (c) Implement the simple substitution estimator (a.k.a., parametric G-computation) using each one of the four regression specifications above. Specifically, for each regression specification for estimating the conditional mean outcome $\mathbb{E}_0(Y|A, W)$, do the following
   - Use `glm` function to estimate $\mathbb{E}_0(Y|A, W)$. Be sure to specify the arguments `family='binomial'` and `data=Obs`.
   - Then use the `predict` function to get the expected outcome for each unit under the intervention $\hat{\mathbb{E}}(Y|A = 1, W_i)$. Be sure to specify the arguments `newdata=txt` and the `type='response'`.
   - Next, use the `predict` function to get the expected outcome for each unit under the control $\hat{\mathbb{E}}(Y|A = 0, W_i)$. Be sure to specify the arguments `newdata=control` and the `type='response'`.
   - Finally, obtain a point estimate of $\Psi(\mathbb{P}_0)$ by substituting the predicted outcomes under the intervention $\hat{\mathbb{E}}(Y|A = 1, W_i)$ and control $\hat{\mathbb{E}}(Y|A = 0, W_i)$ into the G-Computation formula and using the sample proportion to estimate the marginal distribution of baseline covariates:

   $$\hat{\Psi}(\mathbb{P}_n) = \frac{1}{n}\sum_{i=1}^{n}\left[\hat{\mathbb{E}}(Y|A = 1, W_i) - \hat{\mathbb{E}}(Y|A = 0, W_i)\right]$$

   (d) Assign the resulting point estimates 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 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/or come to office hours.

Solution:

```
> # 0. to avoid repeating the same code for a simple substitution in Step3c.
> # Write a function to implement parametric Gcomp
> do.gcomp <- function(Obs, txt, control, reg.formula){
+   reg.fit <- glm(as.formula(reg.formula), family='binomial', data=Obs)
+   predict.txt <- predict(reg.fit, newdata = txt, type='response')
+   predict.control <- predict(reg.fit, newdata = control, type='response')
+   psi.hat <- mean(predict.txt - predict.control)
+   psi.hat
+ }
```

```
> # 1. setting the number of iterations R and number of observations
> R= 500
> n=200
> # 2. matrix for the estimates
> estimates <- matrix(NA, nrow=R, ncol=4)

> # 3. For loop.
> for(r in 1:R){
+
+   # (a) Draw n=200 i.i.d. observations
+   df <- generate.data(n)
+   Obs <- subset(df, select=c(W1,W2,A,Y))
+
+   # (b) Copy the original dataset Obs into two new dataframes txt and control.
+   txt<- control <- Obs
+   # set A=1 in the txt dataframe and A=0 in control dataframe
+   txt$A <-1
+   control$A <- 0
+
+   # (c) Implementing a simple substitution estimator
+   # estimate with the conditional mean function according to parametric model 1
+   psi.hat1 <- do.gcomp(Obs=Obs, txt=txt, control=control,
+                        reg.formula=(Y~A) )
+
+   # estimate with the conditional mean function according to parametric model 2
+   psi.hat2 <- do.gcomp(Obs=Obs, txt=txt, control=control,
+                        reg.formula=(Y ~ A + W1) )
+
+   # estimate with the conditional mean function according to parametric model 3
+   psi.hat3 <- do.gcomp(Obs=Obs, txt=txt, control=control,
+                        reg.formula=(Y ~ A + W2) )
+
+   # estimate with the conditional mean function according to parametric model 4
+   psi.hat4 <- do.gcomp(Obs=Obs, txt=txt, control=control,
+                        reg.formula=(Y ~ A + W1 + W2 + A:W1 + A:W2) )
+
+   #(d) save the resulting estimates
+   estimates[r,] <- c(psi.hat1, psi.hat2, psi.hat3, psi.hat4)
+ }
> estimates <- data.frame(estimates)
> colnames(estimates) <- c('psi.hat1', 'psi.hat2', 'psi.hat3', 'psi.hat4')
> summary(estimates)

    psi.hat1         psi.hat2         psi.hat3         psi.hat4
 Min.   :0.4979   Min.   :0.4253   Min.   :0.3146   Min.   :0.2539
 1st Qu.:0.6126   1st Qu.:0.5835   1st Qu.:0.5238   1st Qu.:0.4513
 Median :0.6547   Median :0.6263   Median :0.5673   Median :0.5079
 Mean   :0.6505   Mean   :0.6228   Mean   :0.5654   Mean   :0.5060
 3rd Qu.:0.6922   3rd Qu.:0.6653   3rd Qu.:0.6135   3rd Qu.:0.5587
 Max.   :0.7849   Max.   :0.7821   Max.   :0.7385   Max.   :0.7336
```

6 Performance of the estimators.

The true value of $\Psi(\mathbb{P}_0)$ is 50.7%.

1. **What is the average point estimate from each?**
2. **Estimate the bias of each estimator.** For each estimator, average the difference between point estimate ψ_n and the truth ψ_0.
3. **Estimate the variance of each estimator.**
4. **Estimate the mean squared error of each estimator.**
5. **Briefly comment on the performance of the estimators in this simulation setting. Which estimator has the lowest MSE over the $R = 500$ iterations? Are you surprised?**

Solution:

```
> # estimatorEval: input a vector of estimates PsiHat and the true parameter value
> # returns a vector with the mean, bias, variance and MSE
> estimatorEval<- function(PsiHat, PsiTrue){
+
+   data.frame(cbind(
+       mean = mean(PsiHat),
+       bias = mean(PsiHat - PsiTrue),
+       var = var(PsiHat),
+       MSE = mean( (PsiHat - PsiTrue)^2)
+   ))
+ }

> # creating a dataframe of a matrix of the estimator's performance
> # run the estimatorEval function on each estimator.
> # the $ is getting a column in data frame estimates
> Psi.P0 <- .507
> performance <- data.frame (rbind(
+   est.1 <- estimatorEval(PsiHat= estimates$psi.hat1, PsiTrue=Psi.P0),
+   est.2 <- estimatorEval(PsiHat= estimates$psi.hat2, PsiTrue=Psi.P0),
+   est.3 <- estimatorEval(PsiHat= estimates$psi.hat3, PsiTrue=Psi.P0),
+   est.4<- estimatorEval(PsiHat= estimates$psi.hat4, PsiTrue=Psi.P0)
+ ))
> round(performance, 3)

   mean   bias   var   MSE
1 0.651  0.144 0.003 0.024
2 0.623  0.116 0.004 0.017
3 0.565  0.058 0.005 0.008
4 0.506 -0.001 0.006 0.006
```

6. The simple substitution estimator using the correctly specified parametric model (Estimator #4) to estimate $\mathbb{E}_0(Y|A,W)$ exhibited the lowest bias and mean squared error. It was more variable than the other estimators. The G-Computation substitution estimator using an unadjusted model for $\mathbb{E}_0(Y|A,W)$ (Estimator #1) had the highest bias and mean squared error. This was not surprising, since the estimator does not control for measured confounders $W1$ and $W2$.

MORAL OF THE STORY: While the simple substitution estimator based on the G-Computation formula is easy to understand, it relies on correctly specifying the conditional mean of the outcome, given the exposure and the covariates.

7 Identifying the mean counterfactual outcome under a dynamic intervention

This section is required, but will be grade leniently. The goal is to improve your understanding of why the backdoor criterion allows us to identify our causal parameter. This problem considers dynamic treatment rules, but the same general arguments also give identifiability for static treatment rules.

Suppose the investigators are also interested in the population mean outcome if, possibly contrary-to-fact, the following dynamic treatment rule d were implemented

$$\begin{aligned} d(W2) &= \mathbb{I}(W2 = 1) \\ &= \begin{cases} 1, & \text{if the child has access to health care} \\ 0, & \text{otherwise.} \end{cases} \end{aligned}$$

That is, the investigators are interested in learning about the causal parameter $\Psi_d^*(\mathbb{P}^*) = \mathbb{E}^*[Y_d]$ where Y_d denotes the counterfactual survival status under this dynamic regime.

If we assume that $W = (W1, W2)$ satisfies the backdoor criterion for the effect of A on Y, then this will imply a randomization assumption for the rule d:

$$Y_d \perp\!\!\!\perp A \mid W1, W2$$

We will also assume the positivity assumption holds.

The objective of this exercise is to understand why $\Psi_d^*(\mathbb{P}^*) = \Psi_d(\mathbb{P}_0)$ under the randomization assumption, where

$$\Psi_d(\mathbb{P}_0) = \sum_{w1,w2} \mathbb{E}_0 \left[Y | A = d(w2), W1 = w1, W2 = w2\right] \mathbb{P}_0 \left(W1 = w1, W2 = w2\right).$$

The derivation of this equality is given below. You are tasked with justifying each of the equalities in the derivation using both properties of random variables and a translation of those properties to the current data structure. We have that

$$\begin{aligned} \Psi_d^*(\mathbb{P}^*) &= \mathbb{E}^*[Y_d] \\ &= \sum_{w1,w2} \mathbb{E}^*[Y_d | W1 = w1, W2 = w2] \mathbb{P}^*(W1 = w1, W2 = w2) && (1) \\ &= \sum_{w1,w2} \sum_y y \mathbb{P}^* \left(Y_d = y | W1 = w1, W2 = w2\right) \mathbb{P}^*(W1 = w1, W2 = w2) && (\star) \\ &= \sum_{w1,w2} \sum_y y \mathbb{P}^* \left(Y_d = y | A = d(w2), W1 = w1, W2 = w2\right) \mathbb{P}^*(W1 = w1, W2 = w2) && (2) \\ &= \sum_{w1,w2} \sum_y y \mathbb{P}^* \left(Y_d = y | A = d(w2), W1 = w1, W2 = w2\right) \mathbb{P}_0(W1 = w1, W2 = w2) && (3) \\ &= \sum_{w1,w2} \sum_y y \mathbb{P}_0 \left(Y = y | A = d(w2), W1 = w1, W2 = w2\right) \mathbb{P}_0(W1 = w1, W2 = w2) && (4) \\ &= \sum_{w1,w2} \mathbb{E}_0[Y | A = d(w2), W1 = w1, W2 = w2] \mathbb{P}_0(W1 = w1, W2 = w2) && (\star) \\ &= \Psi_d(\mathbb{P}_0), \end{aligned}$$

where each $(\star)$ holds by the definition of conditional expectation. Below we ask you to justify labeled equalities (1) through (4). Note that we have implicitly used the positivity assumption in (2) and all of the subsequent equalities. Positivity ensures the conditional expectations and probabilities make sense – it is impossible to calculate the average outcome in a stratum which does not contain any individuals (i.e., occurs with probability 0) in the target population!

1. Explain why (1) holds using properties of conditional expectations.

2. Explain why (2) holds using properties of conditional expectations and the fact that $Y_d \perp\!\!\!\perp A|W1, W2$. `Note: No need to explain` $Y_d \perp\!\!\!\perp A|W1, W2$ `in the context of the study since you have already discussed the assumptions needed for the backdoor criterion to hold, and the backdoor criterion implies` $Y_d \perp\!\!\!\perp A|W1, W2$.
3. Explain why (3) holds.
4. Explain why (4) holds.

Solution:

1. By the law of total expectation,

$$\begin{aligned}\mathbb{E}^*[Y_d] &= \mathbb{E}^*\Big[\mathbb{E}^*[Y_d|W1, W2]\Big] \\ &= \sum_{w1,w2} \mathbb{E}^*[Y_d|W1 = w1, W2 = w2]\mathbb{P}^*(W1 = w1, W2 = w2).\end{aligned}$$

2. If $Y_d \perp\!\!\!\perp A \mid W1, W2$, then knowing that $A = d(w2)$ does not change the probability that $Y_d = y$ within strata of $W1, W2$. That is,

$$\mathbb{P}^*\left(Y_d = y|W1 = w1, W2 = w2\right) = \mathbb{P}^*\left(Y_d = y|A = d(w2), W1 = w1, W2 = w2\right).$$

3. By our SCM, we know that there exist functions f_{W1}, f_{W2}, f_A, and f_Y and some unmeasured factors $U = (U_{W1}, U_{W2}, U_A, U_Y) \sim \mathbb{P}_U$ such that:

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

   In this study, our endogenous variables are $X = (W1, W2, A, Y)$, and we observe them all $O = (W1, W2, A, Y)$.

   We assume this SCM describes the data generating process under existing conditions and under specific interventions. In other words, the same set of functions f_{W1} and f_{W2} generate the baseline covariates $(W1, W2)$ in the "observed data world" and in the "counterfactual world". Given a realization of the background factors $U = u$, the same values $(w1, w2)$ would be realized. This link allows us to replace $\mathbb{P}^*(W1 = w1, W2 = w2)$ with $\mathbb{P}_0(W1 = w1, W2 = w2)$

4. In the SCM above, $A =$ is the *observed* treatment. However, by the assumed autonomy of the structural equations, the same function f_Y generates the observed outcome Y as well as the counterfactual outcome Y_d under our dynamic intervention. (We assume that changing how the exposure was generated does not change the effect of the realized exposure on the outcome.) Furtherore, within strata of the covariates $(w1, w2)$ and at the level of the exposure of interest $(A = d(w2))$, we have that the observed outcome $Y = y$ equals the counterfactual outcome $Y_d = y$:

$$\begin{aligned}Y &= f_Y(W1 = w1, W2 = w2, A = d(w2), U_Y) \\ Y_d &= f_Y(W1 = w1, W2 = w2, A = d(w2), U_Y)\end{aligned}$$

   where the randomness is through U_Y. Thus, within exposure-covariate strata $Y = Y_d$. Note that some causal frameworks make an explicit assumption known as the *consistency assumption* so that $Y_d = Y$ when the observed exposure A is equal to $d(W2)$, but here this was a direct result of our posited SCM. It follows that

$$\mathbb{P}_0(Y = y|A = d(w2), W1 = w1, W2 = w2) = \mathbb{P}^*(Y_d|A = d(w2), W1 = w1, W2 = w2).$$