R Lab 2 - Identifiability & the Simple Substitution Estimator

Laura B. Balzer

Biostat683 - Intro. to Causal Inference

Goals:

1. Review the steps 1-5 of the Roadmap: (1) specify the causal model, (2) specify the causal question, (3) specify the observed data and its link to the causal model, (4) assess identifiability, and (5) specify a statistical estimand and statistical model.
2. Introduce and implement the simple substitution estimator based on the G-Computation formula.
3. Use simulations to evaluate the properties of estimators.

Next lab:
We will implement the inverse probability of treatment weighted (IPTW) estimator and explore the impact of positivity violations on estimator performance.

Reminder:
This is not an `R` class. However, software is an important bridge between the statistical concepts and implementation.

1 Background Story

"[The Hunger Games] is written in the voice of sixteen-year-old Katniss Everdeen, who lives in a post-apocalyptic world in the country of Panem where the countries of North America once existed. The Capitol, a highly advanced metropolis, holds hegemony over the rest of the nation. The Hunger Games are an annual event in which one boy and one girl aged 12 to 18 from each of the 12 districts surrounding the Capitol are selected by lottery [as 'tributes'] to compete in a televised battle in which only one person can survive." - Source: Wikipedia "The Hunger Games"

Some of the tributes have trained extensively for this tournament. The life experiences of other tributes have resulted in certain abilities/advantages (e.g., strength, tree climbing, markmanship). Prior to the tournament, a committee of judges assigns a score to each the tribute indicating their probability of winning. Once the tournament starts, forming alliances and sponsorship can aid in survival. A lone victor returns to their district and is showered with wealth and other resources.

Suppose we are interested in the effect of forming an alliance on the probability of surviving through the first 24 hours. We have randomly sampled one tribute from each year of the games. Let $W1$ denote the tribute's sex with $W1 = 1$ being male and $W1 = 0$ female. Let $W2$ denote the score from the judges. Let A be an indicator that an alliance is formed, and Y be an indicator of survival through the first 24 hours. Finally, let $W3$ be an indicator of whether the tribute receives aid from sponsors during the tournament. Our goal is to evaluate the effect of forming an alliance on the probability of surviving through the first 24 hours.

This study can be translated into the following directed acyclic graph (DAG) shown in Figure 1.

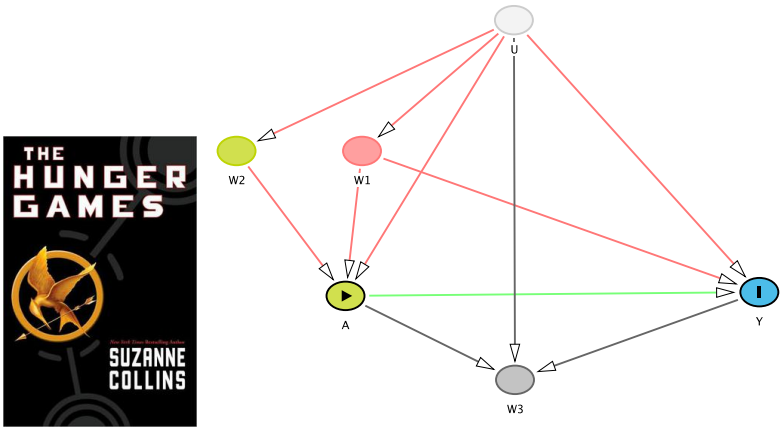

Figure 1: Directed Acyclic Graph for the Hunger Games study.

1. **Translate the DAG into the corresponding structural causal model $\mathcal{M}^*$.**
2. **Are there any exclusion restrictions?**
3. **Are there any restrictions on the distribution of the background variables $\mathbb{P}_U$? In other words, are there any independence assumptions?**
4. **Specify the causal question and parameter.**
5. **Suppose the observed data consist of n independent, identically distributed (i.i.d.) draws of the random variable $O = (W1, W2, A, Y, W3) \sim \mathbb{P}_0$.** Specify the link between the SCM and the observed data. Does the SCM place any restrictions on the statistical model $\mathcal{M}$?

2 Assess identifiability of our causal parameter $\Psi^*(\mathbb{P}^*)$ & Commit to a statistical estimand $\Psi(\mathbb{P}_0)$.

The causal risk difference $\Psi^*(\mathbb{P}^*)$ is not identified under our causal model $\mathcal{M}^*$. A sufficient, but *not* minimal, identifiability assumption is that all of the unmeasured factors are independent. If the all unmeasured factors were conveniently independent, the back-door criterion would hold conditional on $W1$. Equivalently, the counterfactual outcome Y_a would be conditionally independent of the treatment A, given $W1$; this is the randomization assumption and also called "conditional exchangeability". Other possible independence assumptions and the corresponding sufficient sets are given in Figure 2.

For the statistical estimand $\Psi(\mathbb{P}_0)$ to be well-defined, we need additional condition of data support, known as the positivity assumption. There must be a positive probability of each exposure condition within each possible values of the adjustment variable $W1$:

$$\mathbb{P}_0(A = 1|W1 = 1) > 0 \qquad \mathbb{P}_0(A = 1|W1 = 0) > 0$$
$$\mathbb{P}_0(A = 0|W1 = 1) > 0 \qquad \mathbb{P}_0(A = 0|W1 = 0) > 0$$

For this specific example, we need a positive probability of forming and not forming an alliance for both men and women.

Despite lack of identifiability (see the Figure 1), we can still "commit" to an interesting statistical estimand

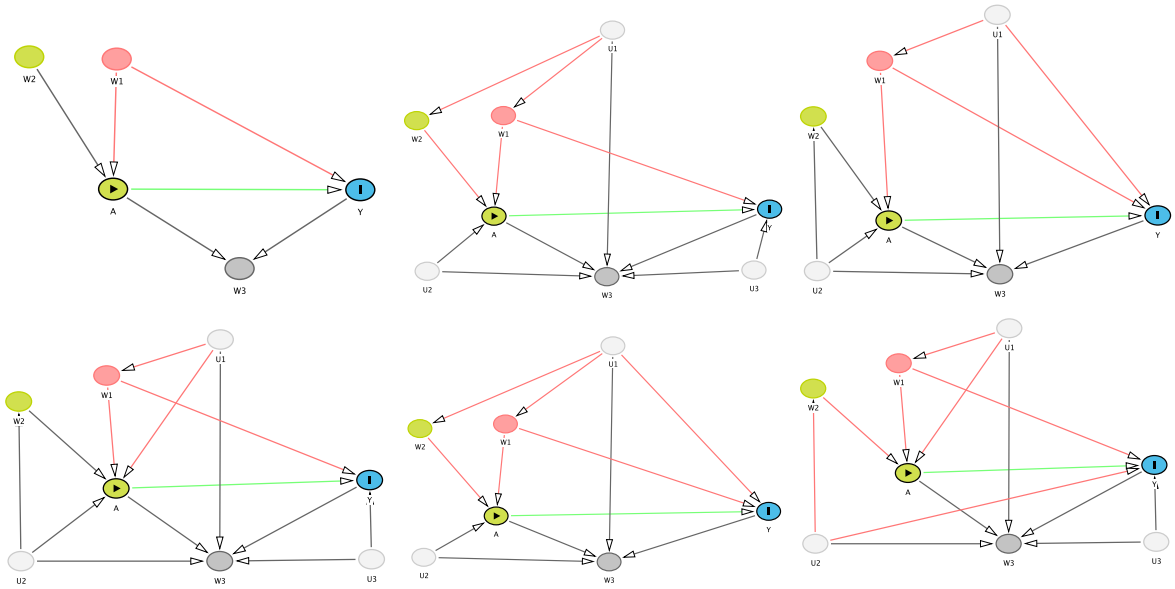

Figure 2: Evaluating the back-door criterion: In the first 4 DAGs, $W1$ alone would satisfy the back-door criterion. In the last 2 DAGs, $W1$ and $W2$ are needed to satisfy the back-door criterion. The needed independence assumptions should be carefully discussed and considered with the help of subject matter experts.

inspired by our scientific/causal question:

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

Formally, the parameter Ψ is a mapping from the statistical model $\mathcal{M}$ to the parameter space $\Psi : \mathcal{M} \to \mathbb{R}$. In other words, Ψ is a function with input as a distribution in $\mathcal{M}$ and output a value in the parameter space (e.g., a number). We have not made any new assumptions during identifiability; therefore, our statistical model $\mathcal{M}$ remains non-parametric.

Note: Alternatively, we could have specified the following as our statistical estimand

$$\Psi^{alt}(\mathbb{P}_0) = \mathbb{E}_0\Big[\mathbb{E}_0(Y|A=1,W1,W2) - \mathbb{E}_0(Y|A=0,W1,W2)\Big]$$

which would equal the causal effect of interest if the independence assumptions in the last 2 DAGs in Figure 2 conveniently held. However, adjusting for both $(W1, W2)$ would require a different and stronger positivity assumption. For simplicity, we will focus on the previous statistical estimand $\Psi(\mathbb{P}_0)$.

3 A specific data generating process

The above SCM is compatible with many possible data generating processes. Recall $\mathcal{M}^*$ is a causal model for the set of possible distributions $\mathbb{P}_{U,X}$ for (U, X). Now, consider the a specific data generating process, where

each of the exogenous nodes U_{X_i} is drawn independently from the following distributions:

$$\begin{aligned} U_{W1} &\sim Uniform(0,1) \\ U_{W2} &\sim Normal(\mu = 1, \sigma^2 = 2^2) \\ U_A &\sim Uniform(0,1) \\ U_Y &\sim Uniform(0,1) \\ U_{W3} &\sim Uniform(0,1) \end{aligned}$$

Given the Us, the endogenous variables are deterministially generated as:

$$\begin{aligned} W1 &= \mathbb{I}\big[U_{W1} < 0.45\big] \\ W2 &= 0.75^* U_{W2} \\ A &= \mathbb{I}\big[U_A < logit^{-1}(-1 + 2.6^* W1 + 0.9^* W2)\big] \\ Y &= \mathbb{I}\big[U_Y < logit^{-1}(-2 + A + 0.7^* W1)\big] \\ W3 &= \mathbb{I}\big[U_{W3} < logit^{-1}(-1 + 1.3^* A + 2.9^* Y)\big] \end{aligned}$$

The $logit^{-1}$ function is the inverse of the logistic function and given by the `plogis` function in `R`:

$$logit(x) = log\left(\frac{x}{1-x}\right) \quad \text{and} \quad logit^{-1}(x) = \frac{1}{1+e^{-x}}$$

We can evaluate the statistical parameter $\Psi(\mathbb{P}_0)$ in closed form:

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

in closed form. In this particular data generating system (one of many compatible with the SCM), $W1$ (sex) is a Bernoulli random variable with mean 0.45:

$$\mathbb{P}_0(W1=1) = \mathbb{E}_0[W1] = 0.45$$

For a given tribute, random error U_{W1} determines whether $W1$ is 1 (male) or 0 (female). Likewise, the binary outcome Y (survival or not) is a Bernoulli random variable with mean given by the $logit^{-1}$ of a function of A and $W1$. Random error U_Y determines whether Y is 1 (survival) or 0 (death). In other words, we know the conditional mean of Y, given A and $W1$:

$$\mathbb{P}_0(Y=1|A,W) = \mathbb{E}_0(Y|A,W) = logit^{-1}(-2 + A + 0.7W1)$$

Plugging these functions into the G-Computation formula and evaluating $\Psi(\mathbb{P}_0)$ in closed form, we have:

$$\begin{aligned} \Psi(\mathbb{P}_0) &= \sum_{w1} \big[\mathbb{E}_0(Y|A=1,W1=w1) - \mathbb{E}_0(Y|A=0,W1=w1)\big]P(W1=w1) \\ &= \big[logit^{-1}(-2+1+0.7*1) - logit^{-1}(-2+0+0.7*1)\big]0.45 \\ &+ \big[logit^{-1}(-2+1+0.7*0) - logit^{-1}(-2+0+0.7*0)\big](1-0.45) \\ &= 0.1775 \end{aligned}$$

```
> # in R the logit^{-1} function is equal to plogis
> Psi.P0<- (plogis(-2+1+0.7*1) - plogis(-2+0+0.7*1) )*0.45 +
+ (plogis(-2+1+0.7*0) - plogis(-2+0+0.7*0))* 0.55
> Psi.P0

[1] 0.1774828
```

We can interpret $\Psi(\mathbb{P}_0)$ as the difference in the sex-specific probability of survival with and without an alliance, averaged with respect to the distribution of sex, is 0.1775. Since the wished-for identifiability assumptions (randomization + positivity) did not hold in our original causal model $\mathcal{M}^*$, we cannot interpret this parameter causally.

4 Translate this data generating process into simulations

1. **Write a function to generate n i.i.d. observations of random variable $O = (W1, W2, A, Y, W3) \sim \mathbb{P}_0$.** As input to this function, use the sample size n. Within this function, simulate the background factors U and evaluate the structural equations F. Recall the $logit^{-1}$ function in R is `plogis`. Give as output of this function, a data frame (`data.frame`) to hold the observed data.
2. **Set the seed to 252 and generate $n = 5000$ observations using your function. Call the output `Obs`. Use the `head` and `summary` functions to examine the output.** The rows are the n repetitions of the data generating process and the columns are the random variables. In other words, the rows are the n participants and the columns are their characteristics.

5 Simple substitution estimator based on the G-Computation formula (a.k.a., parametric G-comp)

In reality, we usually do not know the true distribution of the observed data $\mathbb{P}_0$. Instead, we only have a sample of n i.i.d. observations of O from $\mathbb{P}_0$. An intuitive estimator of the statistical estimand $\Psi(\mathbb{P}_0)$ is the simple substitution estimator based on the G-Computation formula. Briefly, the algorithm estimates the relevant parts of the observed data distribution $\mathbb{P}_0$ and plugs them into the parameter mapping Ψ:

1. Estimate the conditional mean $\mathbb{E}_0(Y|A, W)$ using the observed data as input.
2. Estimate the marginal distribution of baseline covariates $\mathbb{P}_0(W)$ using the observed data as input.
3. Substitute these estimates into the target parameter mapping:

$$\hat{\Psi}(\mathbb{P}_n) = \sum_w \big[\hat{\mathbb{E}}(Y|A = 1, W = w) - \hat{\mathbb{E}}(Y|A = 0, W = w)\big]\hat{\mathbb{P}}(W = w)$$

where $\mathbb{P}_n$ denotes the empirical distribution, which puts weight $1/n$ on each copy $O_i,\ i = 1, \ldots, n$, and where W denotes our adjustment set.
We will always use the sample proportion to estimate the covariate distribution $\mathbb{P}_0(W)$ and therefore can express our simple substitution estimator as

$$\begin{aligned}
\hat{\Psi}(\mathbb{P}_n) &= \sum_w \big[\hat{\mathbb{E}}(Y|A = 1, W = w) - \hat{\mathbb{E}}(Y|A = 0, W = w)\big] \times \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}(W_i = w) \\
&= \frac{1}{n}\sum_{i=1}^{n}\sum_w \big[\hat{\mathbb{E}}(Y|A = 1, W = w) - \hat{\mathbb{E}}(Y|A = 0, W = w)\big] \times \mathbb{I}(W_i = w) \\
&= \frac{1}{n}\sum_{i=1}^{n} \big[\hat{\mathbb{E}}(Y|A = 1, W_i) - \hat{\mathbb{E}}(Y|A = 0, W_i)\big]
\end{aligned}$$

Formally, an estimator $\hat{\Psi}$ is a mapping from the set of possible empirical distributions $\mathbb{P}_n$ to the parameter space ($\mathbb{R}$). In other words, $\hat{\Psi}$ is a function with input as the observed data (a realization of $\mathbb{P}_n$) and output a value in the parameter space (e.g., a number). The estimator should respect the statistical model $\mathcal{M}$, which is non-parametric. In other words, we should not make any unfounded assumptions about the observed data distribution $\mathbb{P}_0$.

5.1 Implementation with parametric regression

Consider the following parametric regression to describe the conditional expectation of the outcome $\mathbb{E}_0(Y|A, W1)$:

$$\mathbb{E}_0(Y|A, W1) = \mathbb{P}_0(Y = 1|A, W1) = logit^{-1}\big(\beta_0 + \beta_1 A + \beta_2 W1 + \beta_3 A \times W1\big)$$

Since both $W1$ and A are binary, this regression is "saturated". Fittings its coefficients and predicting outcomes would be equivalent to using the NPMLE (i.e., taking the mean outcome within in each strata of $(W1, A)$.)

1. **Use the `glm` function to fit the conditional mean function $\mathbb{E}_0(Y|A, W1)$ with logistic regression. Be sure to specify the arguments `family='binomial'` and `data=Obs`.**
   Hint: To get interaction terms, try the formula $Y \sim A + W1 + A * W1$.

2. **Copy the data set `Obs` into two new data frames `txt` and `control`. Then set `A=1` for all units in `txt` and `A=0` for all units in `control`.**
   Hint: Columns of a data frame can be accessed with the `$` operator.

3. **Use the `predict` function to get the expected outcome for each individual i under the intervention $\hat{\mathbb{E}}(Y|A = 1, W1_i)$. Be sure to specify the arguments `newdata=txt` and the `type='response'`.**
   We use `type='response'` to get back predicted probabilities (instead of log-odds ratios).

4. **Use the `predict` function to get the expected outcome for each individual i under the control $\hat{\mathbb{E}}(Y|A = 0, W1_i)$. Be sure to specify the arguments `newdata=control` and the `type='response'`.**

5. **Evaluate the statistical parameter by substituting the predicted outcomes into the G-Computation formula.** As previously discussed, the sample proportion is a non-parametric maximum likelihood estimator of the marginal distribution of $W1$. So we can just take the empirical mean of the difference in the predicted outcomes for each participant:

$$\hat{\Psi}(\mathbb{P}_n) = \frac{1}{n} \sum_{i=1}^{n} \left[\hat{\mathbb{E}}(Y|A = 1, W1_i) - \hat{\mathbb{E}}(Y|A = 0, W1_i) \right]$$

6 Estimate bias, variance and mean squared error (MSE).

Simulations are useful for evaluating the properties of estimators. We will focus on estimating the bias, variance and mean squared error of the simple substitution estimator. Specifically, for $R = 500$ iterations, we will sample $n = 200$ i.i.d. observations from $\mathbb{P}_0$, implement the simple substitution estimator based on the G-Computation formula, and save the resulting estimate ψ_n.

1. **Reset the seed to 252; set `R` to 500 and `n` to 200.**

2. **Create a vector `estimates` of length $R = 500$ to hold the estimated values ψ_n obtained at each iteration.**
   Hint: Use the `rep` function to create a vector of missing values `NA`.

3. **Inside a `for` loop from 1 to $R = 500$, sample n i.i.d. observations of random variable $O = (W1, W2, A, Y, W3)$; implement the simple substitution estimator using the saturated regression model (adjusting for A, $W1$ and their interaction), and save the resulting estimate ψ_n as an entry in the vector `estimates`.**

   Hint: A simple example of a `for` loop is given below. More information on the syntax can be found with `?for`.

```
> # this code creates an empty vector "temp" of length 10
> # in the for loop, the empty values are replaced by 2*index
> temp<- rep(NA, 10)
> for(i in 1:10) {
+   temp[i]<- 2*i
+ }
> temp
```

```
[1]  2  4  6  8 10 12 14 16 18 20
```

4. **What is the average value of the estimates over $R = 500$ repetitions of the data generating process ?**

5. **Estimate the bias of the estimator.** What is the average deviation of the estimate and the truth $\Psi(\mathbb{P}_0)$? Hint: use the `mean` function.

$$Bias\big(\hat{\Psi}(\mathbb{P}_n)\big) = \mathbb{E}_0(\hat{\Psi}(\mathbb{P}_n) - \Psi(\mathbb{P}_0))$$

6. **Estimate the variance of the estimator.** How much do the estimates vary across samples? Hint: use the `var` function.

$$Variance\big(\hat{\Psi}(\mathbb{P}_n)\big) = \mathbb{E}_0\left(\left(\hat{\Psi}(\mathbb{P}_n) - \mathbb{E}_0[\hat{\Psi}(\mathbb{P}_n)]\right)^2\right)$$

7. **Estimate the mean squared error of the estimator.** On average, how far are the estimates from the truth?

$$\begin{aligned} MSE\big(\hat{\Psi}(\mathbb{P}_n)\big) &= \mathbb{E}_0\left(\left(\hat{\Psi}(\mathbb{P}_n) - \Psi(\mathbb{P}_0)\right)^2\right) \\ &= Bias^2 + Variance \end{aligned}$$

7 More practice

Suppose the Capitol (people in charge of the Hunger Games) demand that you estimate the conditional mean outcome, according to following parametric regression model:

$$\mathbb{E}(Y|A, W1, W2, W3) = logit^{-1}\big(\beta_0 + \beta_1 A + \beta_2 W1 + \beta_2 W2 + \beta_3 W3\big)$$

In other words, they believe that conditional probability of survival through the first 24 hours is a linear (on the logit scale) function of the exposure (alliance), all the pre-exposure covariates $(W1, W2)$ and a post-exposure covariate $(W3)$. This "knowledge" changes our SCM $\mathcal{M}^*$, because it restricts the set of allowed functions f_Y. This "knowledge" also changes our statistical model $\mathcal{M}$, because it restricts the allowed conditional distributions for Y given $(A, W1, W2, W3)$.

1. **Does the back-door criterion hold conditional on $W1$, $W2$ and $W3$ (assuming independence of the Us)?**

2. **For $R = 500$ iterations, repeat the above process of sampling $n = 200$ observations, fitting the conditional mean outcome with a main terms logistic model (adjusting now for A, $W1$, $W2$ and $W3$), obtaining the predicted values under $A = 1$ and $A = 0$, and substituting the estimates into the target parameter mapping.** Don't forget to reset the seed.

3. **Compare the bias, variance and mean squared error of the substitution estimators when using the previous saturated model (adjusting for A, $W1$ and their interaction) and the Capitol's parametric model (adjusting for A, $W1$, $W2$, $W3$ with main terms) to estimate the conditional mean outcome $\mathbb{E}_0(Y|A, W)$.**