R Lab 2 - Identifiability & the Simple Substitution Estimator

Introduction 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. Obtain the value the statistical estimand closed form.
3. Obtain the value the statistical estimand with simulations.
4. Introduce and implement the simple substitution estimator based on the G-Computation formula.
5. Use simulations to evaluate the properties of estimators.

Next lab:
Code discrete SuperLearner to select the estimator with the lowest cross-validated risk. Use `R SuperLearner` package to build the best convex combination of candidate algorithms and to evaluate the performance of SuperLearner.

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 engendered 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 his/her likelihood 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 gender with $W1 = 1$ being male and $W1 = 0$ female. Let $W2$ denote the score from the judges. Let A be an indicator of whether an alliance is formed or not, 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:

1. **Translate the DAG into the corresponding structural causal model $\mathcal{M}^F$.**

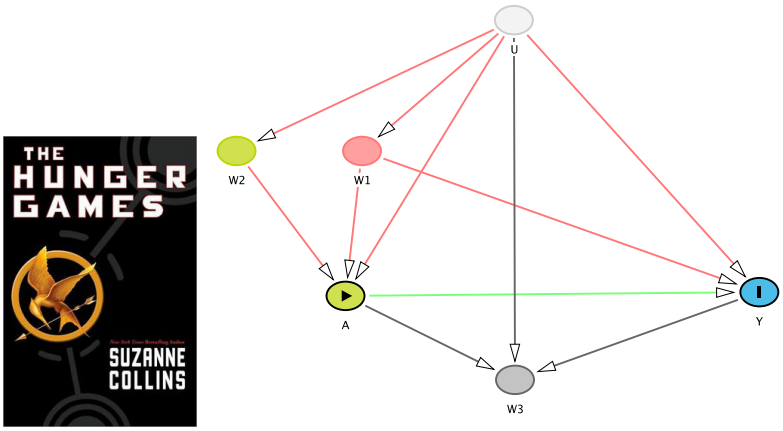


Figure 1: Directed Acyclic Graph for the Hunger Games study.

2. **Are there any exclusion restrictions?**
3. **Are there any restrictions on the distribution of the exogenous variables 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 $O = (W1, W2, A, Y, W3) \sim P_0$. Specify the link between the SCM and the observed data. Does the SCM place any restrictions on the statistical model $\mathcal{M}$?**
6. **Using the backdoor criteria, assess identifiability of $\Psi^F(P_{U,X})$. If not identified, under what assumptions would it be?**
7. **Specify the target parameter of the observed data distribution $\Psi(P_0)$.**

Solution:

1. Endogenous variables: $X = (W1, W2, A, Y, W3)$
   Exogenous variables: $U = (U_{W1}, U_{W2}, U_A, U_Y, U_{W3}) \sim P_U$
   Structural equations F:

$$\begin{aligned} W1 &= f_{W1}(U_{W1}) \\ W2 &= f_{W2}(U_{W2}) \\ A &= f_A(W1, W2, U_A) \\ Y &= f_Y(W1, A, U_Y) \\ W3 &= f_{W3}(A, Y, U_{W3}) \end{aligned}$$

2. We have made lots of exclusion restrictions. Gender $W1$ does not affect the judge's score $W2$. The outcome Y is not affected by the score $W2$. Whether the tribute receives aid from sponsors is not a function of gender $W1$ or the score $W2$.
3. There are no independence assumptions.

4. The target causal parameter is the difference in the counterfactual probability of survival through the first 24 hours, if all tributes formed alliances, and the counterfactual probability of survival, if all tributes did not form alliances:

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

   where Y_a denotes the counterfactual outcome under an intervention to set alliance status $A = a$.

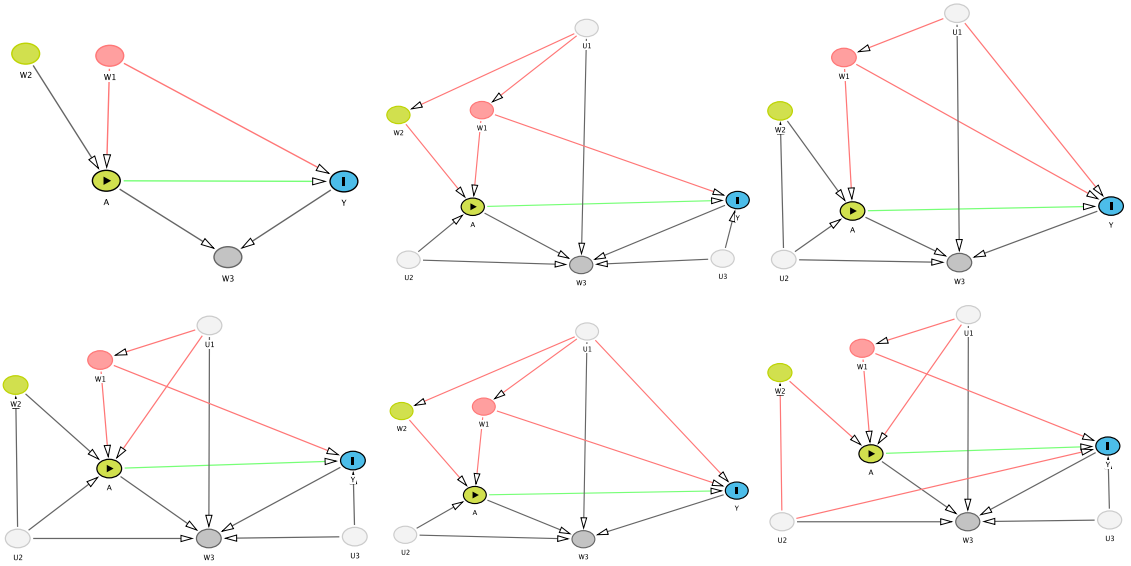
5. We assume the observed data $O = (W1, W2, A, Y, W3)$ were generated by sampling n i.i.d. times from a data generating system compatible with $\mathcal{M}^F$. This provides a link between the causal model $\mathcal{M}^F$ and the observed data O. The distribution of the exogenous variables U and the structural equations F identify the distribution of the endogenous variables X and thus the distribution of the observed data O. We have not placed any restrictions on the statistical model $\mathcal{M}$, which is thereby non-parametric.

6. In the SCM $\mathcal{M}^F$, the target causal quantity is not identified. A sufficient, but *not* minimal, identifiability assumption is that all of the exogenous errors an independent. Other possible independence assumptions and the corresponding sufficient sets are given in Figure 1. We use $\mathcal{M}^{F*}$ to denote the original SCM, augmented with additional assumptions needed for identifiability. We introduce this "working" SCM to keep our real knowledge separate from our wished identifiability assumptions. Under $\mathcal{M}^{F*}$, the backdoor criteria holds conditional on $W1$. Equivalently, the counterfactual outcome Y_a is conditionally independent of the treatment A, given $W1$. This is the randomization assumption.

7. Under the working SCM $\mathcal{M}^{F*}$, the causal risk difference $\Psi^F(P_{U,X})$ is identified using the G-Computation formula:

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

   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). *Note:* For the statistical estimand to be well-defined, we need additional condition of data support, known as the positivity assumption. This will be discussed in detail in the coming weeks.



Solution Fig. 1: Evaluating the backdoor criteria: In the first 4 DAGs, $W1$ alone would satisfy the backdoor criteria. In the last 2 DAGs, $W1$ and $W2$ are needed to satisfy the backdoor criteria. The needed independence assumptions should be carefully discussed and considered with the help of subject matter experts. For simplicity, we are assuming a working SCM $\mathcal{M}^{F*}$ where all the exogenous variables U are independent.

2 A specific data generating process

The above SCM is compatible with many possible data generating processes. Recall $\mathcal{M}^F$ is a causal model for the set of possible distributions $P_{U,X}$ 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 U's, the endogenous variables are deterministially generated as:

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

The *expit* function is the inverse of the logistic function:

$$\begin{aligned} logit(x) &= log\left(\frac{x}{1-x}\right) \\ expit(x) &= \frac{1}{1+e^{-x}} \end{aligned}$$

1. **Evaluate $\Psi(P_0)$ in closed form.** Hint: plug in the necessary functions into the G-Computation formula.
2. **Interpret $\Psi(P_0)$.**

Solution:

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

   $$P_0(W1 = 1) = E_0[W1] = 0.45$$

   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 *expit* 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 $W1$:

   $$P_0(Y = 1|A, W) = E_0(Y|A, W) = expit(-2 + A + 0.7W1)$$

Plugging these functions into the G-Computation formula and evaluating $\Psi(P_0)$ in closed form, we have:

$$\begin{aligned}\Psi(P_0) &= \sum_{w1} \big[E_0(Y|A=1, W1=w1) - E_0(Y|A=0, W1=w1)\big] P(W1=w1) \\ &= \big[expit(-2+1+0.7*1) - expit(-2+0+0.7*1)\big] 0.45 \\ &+ \big[expit(-2+1+0.7*0) - expit(-2+0+0.7*0)\big](1-0.45) \\ &= 0.1775\end{aligned}$$

```
> # in R the expit 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
```

2. The difference in the gender-specific probability of survival under the intervention and under the control, averaged with respect to the distribution of gender is 0.1775. Under the randomization assumption, $\Psi(P_0)$ can be interpreted as the causal risk difference: the probability of survival through the first 24 hours would be 17.75% higher under an intervention to form an alliance than under an intervention prevent an alliance.

3 Translate this data generating process into simulations

1. **First set the seed to 252.**
2. **Set the number of draws** $n = 5000$**.**
3. **Sample** n **i.i.d. observations of random variable** $O = (W1, W2, A, Y, W3) \sim P_0$**.** In other words, simulate the background factors U and evaluate the structural equations F. The *expit* function in R is `plogis`.
4. **Create a data frame (`data.frame`), named `Obs`, to hold these values.** The rows are the n repetitions of the experiment and the columns are the random variables. In other words, the rows are the n tributes and the columns are their characteristics. **Use the `head` and `summary` functions to get a better understanding of the data generating experiment.**

Solution:

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


> # 2. set the number of draws
> n <- 5000


> # 3. Draw n i.i.d. observations
> U.W1<- runif(n, min=0, max=1)
> U.W2<- rnorm(n, mean=1, sd=2)
> U.A <- runif(n, min=0, max=1)
```

```
> U.Y <- runif(n, min=0, max=1)
> U.W3<- runif(n, min=0, max=1)
> #
> W1<- as.numeric( U.W1 < 0.45)
> W2<- 0.75*U.W2
> A <- as.numeric( U.A < plogis(-1+2.6*W1+0.9*W2))
> Y <- as.numeric( U.Y < plogis(-2+A+0.7*W1))
> W3<- as.numeric( U.W3 < plogis(-1 + 1.3*A + 2.9*Y))


> # 4. dataframe O for the observed data.
> Obs<- data.frame(W1, W2, A, Y, W3)
> head(Obs)


  W1         W2 A Y W3
1  0  0.9012694 1 0  1
2  0 -0.2273375 0 0  1
3  1  0.5045215 1 1  1
4  0  1.2643669 1 0  1
5  0  5.3133210 1 1  1
6  1  1.2645811 1 0  0


> summary(Obs)


       W1               W2                A                Y
 Min.   :0.0000   Min.   :-4.0200   Min.   :0.0000   Min.   :0.0000
 1st Qu.:0.0000   1st Qu.:-0.2676   1st Qu.:0.0000   1st Qu.:0.0000
 Median :0.0000   Median : 0.7765   Median :1.0000   Median :0.0000
 Mean   :0.4362   Mean   : 0.7356   Mean   :0.6168   Mean   :0.2712
 3rd Qu.:1.0000   3rd Qu.: 1.7261   3rd Qu.:1.0000   3rd Qu.:1.0000
 Max.   :1.0000   Max.   : 7.5540   Max.   :1.0000   Max.   :1.0000
       W3
 Min.   :0.0000
 1st Qu.:0.0000
 Median :1.0000
 Mean   :0.5692
 3rd Qu.:1.0000
 Max.   :1.0000
```

It is worth re-iterating that R generates the exogenous input U by calling a pseudorandom number generator to simulate a $Uniform(0,1)$ variable and then transformming it to correspond with a draw from the specified distribution. We went through an equivalent process above when we drew U_{W1} from a uniform distribution and then using an indicator function to generate a Bernoulli random variable with probability p. In many cases, we can simplify the R code and directly simulate the endogenous variables as follows:

```
> W1 <-rbinom(n, size=1, prob=0.45)
> W2 <- 0.75*rnorm(n, mean=1, sd=2)
> A <- rbinom(n, size=1, prob=plogis(-1+2.6*W1+0.9*W2))
> Y <- rbinom(n, size=1, prob=plogis(-2+A+0.7*W1))
> W3<- rbinom(n, size=1, prob=plogis(-1+1.3*A+2.9*Y))
```

4 Simple substitution estimator based on the G-Computation formula

In the Section 2, we used our knowledge of the true distribution of the observed data P_0 to obtain the value of the target parameter. Specifically, we plugged in the true conditional mean $E_0(Y|A,W)$ and the marginal distribution $P_0(W)$ into the G-computation formula:

$$\Psi(P_0) = E_0\big[E_0(Y|A=1,W) - E_0(Y|A=0,W)\big]$$

where W represents the covariates that satisfy the backdoor criteria for the effect of A on Y. In our example, $W1$ satisfies the backdoor criteria under the working model $\mathcal{M}^{F*}$.

In reality, 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 from P_0. An intuitive estimator of the statistical estimand $\Psi(P_0)$ is the simple substitution estimator based on the G-Computation formula. Briefly, the algorithm estimates the relevant parts of the observed data distribution P_0 and plugs them into the parameter mapping Ψ:
(A.) Estimate the conditional mean $E_0(Y|A,W)$ using the observed data as input.
(B.) Estimate the marginal distribution of baseline covariates $P_0(W)$ using the observed data as input.
(C.) Substitute these estimates into the target parameter mapping:

$$\hat{\Psi}(P_n) = \sum_{w1} \big[\hat{E}(Y|A=1,W=w) - \hat{E}(Y|A=0,W=w)\big]\hat{P}(W=w)$$

where P_n denotes the empirical distribution, which puts weight $1/n$ on each copy $O_i,\ i=1,\ldots,n$.

Formally, an estimator $\hat{\Psi}$ is a mapping from the set of possible empirical distributions 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 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 P_0. In lecture, we will go over this estimator and its properties in detail.

4.1 Implementation with the NPMLE

1. **Estimate the conditional mean function with the non-parametric maximum likelihood estimator (NPMLE). Create strata of each possible value of $(A, W1)$ and take the empirical mean of Y in each strata.** This is equivalent to fitting a saturated regression model.

   Hint: The following code creates a vector of the outcomes among unexposed ($A=0$) females ($W1=0$). The NPMLE for the conditional probability of survival for this subgroup is the empirical mean of the resulting vector:

   ```
   > # outcomes among unexposed females
   > Y.a0w0<-  Y[W1==0 & A==0]
   > meanY.a0w0 <- mean(Y.a0w0)
   > meanY.a0w0

   [1] 0.1207116
   ```

   In words, the observed probability of survival among females, who did not make an alliance, is $\hat{E}(Y|A=0,W1=0) = 12.07\%$.

2. **Estimate the marginal distribution $P_0(W1=w1)$ with the sample proportion**:

   $$\hat{P}(W1=w1) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{I}(W1_i = w1)$$

   Again, this non-parametric estimator does not place any restrictions on the statistical model.

3. **Substitute these estimates into the parameter mapping**:

$$\begin{aligned}\hat{\Psi}(P_n) =& \big[\hat{E}(Y|A=1,W1=1) - \hat{E}(Y|A=0,W1=1)\big]\hat{P}(W1=1) \\ &+ \big[\hat{E}(Y|A=1,W1=0) - \hat{E}(Y|A=0,W1=0)\big]\hat{P}(W1=0)\end{aligned}$$

Solution:

```
> # 1. Estimating E(Y|A,W) with NPMLE
>
> # mean outcome among unexposed females
> meanY.a0w0<- mean(Y[W1==0 & A==0])
> meanY.a0w0

[1] 0.1207116

> # mean outcome among exposed females
> meanY.a1w0<- mean(Y[W1==0 & A==1])
> meanY.a1w0

[1] 0.2586345

> # mean outcome among unexposed males
> meanY.a0w1<- mean(Y[W1==1 & A==0])
> meanY.a0w1

[1] 0.2017544

> # mean outcome among exposed males
> meanY.a1w1<- mean(Y[W1==1 & A==1])
> meanY.a1w1

[1] 0.4214247

> # 2 Estimating the marginal distribution P(W1) with the NPMLE
> # prob W1=1 is simple the mean of W1
> mean(W1)

[1] 0.4362

> # 3. Substitute the estimates into the target parameter mapping (G-comp formula)
> Psi.hat<- (meanY.a1w1 - meanY.a0w1)*mean(W1) + (meanY.a1w0 - meanY.a0w0)*(1-mean(W1))
> Psi.hat

[1] 0.1735812
```

4.2 Implementation with parametric regression

In the previous subsection, we estimated the conditional risk $E_0(Y|A, W1)$ with the empirical mean outcome Y in strata of A and $W1$. This is equivalent to fitting a saturated parametric model for the conditional mean:

$$E_0(Y|A, W1) = P_0(Y = 1|A, W1) = expit\big(\beta_0 + \beta_1 A + \beta_2 W1 + \beta_3 AW1\big)$$

1. **To gain familiarity with R and the simple substitution estimator, use the `glm` function to fit the conditional mean function $E_0(Y|A, W1)$ with logistic regression. Be sure to specify the arguments `family='binomial'` and `data=Obs`.**
   Hint: To get interaction terms, try the formula $Y \sim A * W1$.

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. **Now use the `predict` function to get the expected outcomes for each individual under the intervention $\hat{E}(Y|A = 1, W1)$. Be sure to specify the arguments `newdata=txt` and the `type='response'`.**

4. **Now use the `predict` function to get the expected outcomes for each individual under the control $\hat{E}(Y|A = 0, W1)$. Be sure to specify the arguments `newdata=control` and the `type='response'`.**

5. **Evaluate the statistical parameter by substituting the predicted mean outcomes under the treatment and under the control into the G-Computation formula.** 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 subject:

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

Solution:

```
> #1.  Estimate the conditional mean of Y given the treatment A and W1
> reg.model<- glm(Y ~ A * W1, family='binomial', data=Obs)
> reg.model


Call:  glm(formula = Y ~ A * W1, family = "binomial", data = Obs)

Coefficients:
(Intercept)            A           W1         A:W1
    -1.9857       0.9326       0.6103       0.1258

Degrees of Freedom: 4999 Total (i.e. Null);  4996 Residual
Null Deviance:            5844
Residual Deviance: 5431         AIC: 5439


> #3. Copy the original dataset O into two new dataframes txtPop and controlPop.
> txt<- control <- Obs
> # set A=1 in the txt dataframe and A=0 in control dataframe
> txt$A <-1
> control$A <- 0
```

```
> # 4 predict the mean outcome for each individual in the sample under the treatment
> Y1.predict<- predict(reg.model, newdata = txt, type='response')


> # 5.  predict the mean outcome for each individual in the sample under the control
> Y0.predict<- predict(reg.model, newdata = control, type='response')
> #
> head(cbind(W1,W2,A,Y, W3, Y1.predict, Y0.predict))


  W1         W2 A Y W3 Y1.predict Y0.predict
1  0  0.9012694 1 0  1  0.2586345  0.1207116
2  0 -0.2273375 0 0  1  0.2586345  0.1207116
3  1  0.5045215 1 1  1  0.4214247  0.2017544
4  0  1.2643669 1 0  1  0.2586345  0.1207116
5  0  5.3133210 1 1  1  0.2586345  0.1207116
6  1  1.2645811 1 0  0  0.4214247  0.2017544


> tail(cbind(W1,W2,A,Y, W3, Y1.predict, Y0.predict))


     W1         W2 A Y W3 Y1.predict Y0.predict
4995  1  2.7779075 1 0  1  0.4214247  0.2017544
4996  0  1.1455181 0 0  0  0.2586345  0.1207116
4997  0  2.0852541 0 0  0  0.2586345  0.1207116
4998  0  0.5176812 1 1  1  0.2586345  0.1207116
4999  0 -1.5372069 0 0  1  0.2586345  0.1207116
5000  1  0.2457420 1 1  1  0.4214247  0.2017544


> # 6. take the mean of the predicted outcomes over the distribution of W1
> mean(Y1.predict - Y0.predict)


[1] 0.1735812
```

5 Estimate the bias, variance and mean squared error (MSE) of the substitution estimator.

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 P_0, implement the simple substitution estimator based on the G-Computation formula, and save the resulting estimate ψ_n.

1. **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 and $W1$), 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 of $R = 500$ trials?**

5. **Estimate the bias of the estimator.** What is the average deviation of the estimate and the truth $\Psi(P_0)$? Hint: use the `mean` function.

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

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

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

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

$$\begin{aligned} MSE\big(\hat{\Psi}(P_n)\big) &= E_0\left(\left(\hat{\Psi}(P_n) - \Psi(P_0)\right)^2\right) \\ &= Bias^2 + Variance \end{aligned}$$

Solution:

```
> # 1. setting the number of iterations and the sample size
> R <- 500
> n <- 200


> # 2. create a vector for the estimates
> estimates<- rep(NA,R)


> # 3. repeat the data generating experiment and estimation algorithm R times
> for(i in 1:R){
+
+   # 1. simulate the sample of n observations
+   W1 <-rbinom(n, size=1, prob=0.45)
+   W2 <- 0.75*rnorm(n, mean=1, sd=2)
+   A <- rbinom(n, size=1, prob=plogis(-1+2.6*W1+0.9*W2))
+   Y <- rbinom(n, size=1, prob=plogis(-2+A+0.7*W1))
+   W3<- rbinom(n, size=1, prob=plogis(-1+1.3*A+2.9*Y))
```

```
+   Obs<- data.frame(W1,W2,A,Y,W3)
+
+   #2  Estimate the conditional mean of Y given the treatment A and W1
+   reg.model<- glm(Y ~ A*W1, family='binomial', data=Obs)
+
+   #3. Copy the original dataset O into two new dataframes txtPop and controlPop.
+   txt<- control <- Obs
+   # set A=1 in the txt dataframe and A=0 in control dataframe
+   txt$A <-1
+   control$A <- 0
+
+   # 4 predict the outcome for each individual in the sample under the treatment
+   Y1.predict<- predict(reg.model, newdata = txt, type='response')
+
+   # 5 predict the outcome for each individual in the sample under the control
+   Y0.predict<- predict(reg.model, newdata = control, type='response')
+
+   # 6.  take the mean of the predicted outcomes over the distribution of W1
+   estimates[i]<- mean(Y1.predict - Y0.predict)
+ }


> # 3-6  average value, bias, variance, and MSE of the estimator
> meanEst<- mean(estimates)
> meanEst


[1] 0.1805442


> bias<- mean(estimates - Psi.P0)
> bias


[1] 0.003061437


> var<- var(estimates)
> var


[1] 0.004863207


> mse<- mean( (estimates-Psi.P0)^2)
> mse


[1] 0.004862853


> # check that mse=bias^2 + var
```

Over $R = 500$ repetitions the average value of the estimand is 0.1805. The true value of the statistical estimand was 0.1775. The estimator has very low bias of 0.003 and a variance of 0.0049. Indeed, its mean squared error is 0.0049 and is dominated by the variance.

6 More practice

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

$$E_0(Y|A, W1, W2, W3) = expit\big(\beta_0 + \beta_1 A + \beta_2 W1 + \beta_2 W2 + \beta_3 W3\big)$$

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 intervention (alliance), all the pre-exposure covariates ($W1, W2$) and a post-exposure covariate ($W3$). This "knowledge" changes our SCM $\mathcal{M}^F$, 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 backdoor criteria hold conditional on $W1$, $W2$ and $W3$ (assuming independence of the errors)?**
2. **For $R = 500$ iterations, repeat the above process of sampling $n = 200$ observations, fitting the conditional mean outcome with a 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.**
3. **Compare the bias, variance and mean squared error of the substitution estimators when using a saturated model (equivalent to the NPMLE) and a misspecified parametric model to estimate the conditional mean $E_0(Y|A, W)$.**

Solution:

1. No the backdoor criteria does not hold conditional on $W1$, $W2$ and $W3$. $W3$ is a collider of the intervention A and the outcome Y. By adjusting for $W3$, we are inducing an association between A and Y. The resulting estimate will not equal the causal risk difference.

```
> estimates.miss <- rep(NA,R)
> for(i in 1:R){
+   # simulate the sample of n=200 observations
+   W1 <-rbinom(n, size=1, prob=0.45)
+   W2 <- 0.75*rnorm(n, mean=1, sd=2)
+   A <- rbinom(n, size=1, prob=plogis(-1+2.6*W1+0.9*W2))
+   Y <- rbinom(n, size=1, prob=plogis(-2+A+0.7*W1))
+   W3<- rbinom(n, size=1, prob=plogis(-1+1.3*A+2.9*Y))
+   Obs<- data.frame(W1,W2,A,Y,W3)
+
+   #2  Estimate the conditional mean of Y given the treatment A, W1, W2,W3
+   miss.model<- glm(Y ~ A + W1 +W2+W3, family='binomial', data=Obs)
+
+   #3. Copy the original dataset O into two new dataframes txtPop and controlPop.
+   txt<- control <- Obs
+   # set A=1 in the txt dataframe and A=0 in control dataframe
+   txt$A <-1
+   control$A <- 0
+
+   # 4 predict the outcome for each individual in the sample under the treatment
+   Y1.predict<- predict(miss.model, newdata = txt, type='response')
```

```
+
+   # 5. predict the outcome for each individual in the sample under the control
+   Y0.predict<- predict(miss.model, newdata = control, type='response')
+
+   # 6.  take the mean of the predicted outcomes over the distribution of W1
+   estimates.miss[i]<- mean(Y1.predict - Y0.predict)
+ }


> # Evaluating the estimator
> meanEst.miss<- mean(estimates.miss)
> bias.miss<- mean(estimates.miss - Psi.P0)
> var.miss<- var(estimates.miss)
> mse.miss<- mean( (estimates.miss-Psi.P0)^2)


> # 3. compare bias, variance, mse of substitution estimators
> estComparison<- data.frame(rbind( c(meanEst, bias, var, mse),
+     c(meanEst.miss, bias.miss, var.miss, mse.miss) ) )
> rownames(estComparison)<- c('Correctly', 'Misspecified')
> colnames(estComparison)<- c('Mean estimate','Bias', 'Var', 'MSE')
> signif(estComparison, 2)


             Mean estimate    Bias    Var    MSE
Correctly             0.18  0.0031 0.0049 0.0049
Misspecified          0.05 -0.1300 0.0059 0.0220
```

When a misspecified model is used to estimate the conditional mean outcome given the intervention and covariates, the resulting substitution estimator is biased. This is unsurprising as $W3$ (receiving sponsorship) is a collider of the intervention (alliance or not) and the outcome (survival through the first 24 hours). Indeed, the average estimate using a misspecified regression model is 0.05. The absolute bias over $R = 500$ repetitions of the experiment is over 40 times higher when unfounded assumptions are placed on the statistical model $\mathcal{M}$.