R Homework Two

Katherine Wolf
Introduction to Causal Inference (PH252D)
March 30, 2020

1 Time to prevent child malnutrition in Sahel

2 A specific data generating process

2.1 Evaluate the positivity assumption in closed form for this data generating process.

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 ($W2$), i.e., all of the following must hold:

$$
\begin{aligned}
0 < \mathbb{P}_0(A = 1 | W1 = 1, W2 = 1) \\
0 < \mathbb{P}_0(A = 0 | W1 = 1, W2 = 1) \\
0 < \mathbb{P}_0(A = 1 | W1 = 1, W2 = 0) \\
0 < \mathbb{P}_0(A = 0 | W1 = 1, W2 = 0) \\
0 < \mathbb{P}_0(A = 1 | W1 = 0, W2 = 1) \\
0 < \mathbb{P}_0(A = 0 | W1 = 0, W2 = 1) \\
0 < \mathbb{P}_0(A = 1 | W1 = 0, W2 = 0) \\
0 < \mathbb{P}_0(A = 0 | W1 = 0, W2 = 0)
\end{aligned}
$$

This data generating process specifies

- That the exogenous factors influencing the value of A are generated as $U_A \sim Uniform(0, 1)$, and
- That given the exogenous factors U_A, the value of A is derministically generated as

$$A = \mathbb{I}[U_A < logit^{-1}(-0.5 + W1 - 1.5 * W2)],$$

  where $A = 1$ if the value of U_A is less than the expression on the right of the inequality sign and $A = 0$ otherwise.

Since the $logit^{-1}$ function is bounded between 0 and 1, and since $U_A \sim Uniform(0, 1)$, i.e., equally likely to be any number from 0 to 1, the probability that $A = 1$ conditional on the covariates $W1$ and $W2$ in this data generating system is then the value of $logit^{-1}(-0.5 + W1 - 1.5 * W2)$, i.e.,

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

Since 0 is the only other possible value for A, and since the total probability must sum to 1, the probability of $A = 0$ conditional on the covariates $W1$ and $W2$ in this data generating system is then the one minus that value, i.e.,

$$\mathbb{P}_0(A = 0 | W1, W2) = 1 - logit^{-1}(-0.5 + W1 - 1.5 * W2).$$

To evaluate the positivity assumption in closed form for this data-generating process, then, we can plug the four possible combinations of $W1$ and $W2$ into the equations above. The positivity assumption is satisfied if all the equations generate a number above 0 for all four possible covariate combinations.

- For $W1 = 1, W2 = 1$:

$$\mathbb{P}_0(A = 1|W1 = 1, W2 = 1) = logit^{-1}(-0.5 + 1 - 1.5 * 1) = 0.2689414$$
$$\mathbb{P}_0(A = 0|W1 = 1, W2 = 1) = 1 - logit^{-1}(-0.5 + 1 - 1.5 * 1) = 0.7310586$$

- For $W1 = 1, W2 = 0$:

$$\mathbb{P}_0(A = 1|W1 = 1, W2 = 0) = logit^{-1}(-0.5 + 1 - 1.5 * 0) = 0.6224593$$
$$\mathbb{P}_0(A = 0|W1 = 1, W2 = 0) = 1 - logit^{-1}(-0.5 + 1 - 1.5 * 0) = 0.3775407$$

- For $W1 = 0, W2 = 1$:

$$\mathbb{P}_0(A = 1|W1 = 0, W2 = 1) = logit^{-1}(-0.5 + 0 - 1.5 * 1) = 0.1192029$$
$$\mathbb{P}_0(A = 0|W1 = 0, W2 = 1) = 1 - logit^{-1}(-0.5 + 0 - 1.5 * 1) = 0.8807971$$

- For $W1 = 0, W2 = 0$:

$$\mathbb{P}_0(A = 1|W1 = 0, W2 = 0) = logit^{-1}(-0.5 + 0 - 1.5 * 0) = 0.3775407$$
$$\mathbb{P}_0(A = 0|W1 = 0, W2 = 0) = 1 - logit^{-1}(-0.5 + 0 - 1.5 * 0) = 0.6224593$$

Since $0 < \mathbb{P}_0(A = 1|W1, W2)$ and $0 < \mathbb{P}_0(A = 0|W1, W2)$ for all four possible combinations of $W1$ and $W2$, the positivity assumption is satisfied.

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

The target causal parameter $\Psi^F(\mathbb{P}_{U,X})$ is the difference between the counterfactual probability of survival (of $Y = 1$) if all children receive the combination prevention package and the counterfactual probability of survival if all children do not receive the package. (Because Y is a Bernoulli variable, its probabilty of equaling one, $\mathbb{P}(Y = 1)$, equals its expectation $\mathbb{E}(Y)$.) Formally,

$$\Psi^F(\mathbb{P}_{U,X}) = \mathbb{P}_{U,X}(Y_1 = 1) - \mathbb{P}_{U,X}(Y_0 = 1) = \mathbb{E}_{U,X}(Y_1) - \mathbb{E}_{U,X}(Y_0)$$

Under the assumptions of the working structural causal model $\mathcal{M}^{\mathcal{F}^*}$, the target causal parameter $\Psi^{F^*}(\mathbb{P}_{U,X}) = \mathbb{E}_{U,X}(Y_1) - \mathbb{E}_{U,X}(Y_0)$ is identified by statistical estimand $\Psi(\mathbb{P}_0)$ using the G-computation formula:

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

This specific data generating process specifies

- That the exogenous factors influencing the value of Y are generated as $U_Y \sim Uniform(0, 1)$ and
- That, given the exogenous factors U_Y, the exposure A, and the covariates $W1$ and $W2$, the value of Y is derministically generated as

$$Y = \mathbb{I}[U_Y < logit^{-1}(-0.75 + W1 - 2W2 + 2.5 * A + A * W1)],$$

  where $Y = 1$ if the value of U_Y is less than the expression on the right of the inequality sign and $Y = 0$ otherwise.

Since the $logit^{-1}$ function is bounded between 0 and 1, and since $U_Y \sim Uniform(0, 1)$, i.e., equally likely to be any number from 0 to 1, the probability that $Y = 1$ conditional on the exposure A and covariates $W1$ and $W2$ in this data generating process is then the value of $logit^{-1}(-0.75 + W1 - 2W2 + 2.5A + A * W1)$, i.e.,

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

Since Y is a Bernoulli random variable, its (conditional) probability of equaling one $\mathbb{P}_0(Y = 1|A, W1, W2)$ equals its expectation $\mathbb{E}_0(Y|A, W1, W2)$, i.e.,

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

Similarly, for W1 and W2, this specific data generating process specifies

- That the exogenous factors influencing the value of W1 and W2 are generated as $U_{W1} \sim Uniform(0, 1)$ and $U_{W2} \sim Uniform(0, 1)$, respectively, and
- That, given the exogenous factors U_{W1} and U_{W2}, respectively, the values of $W1$ and $W2$ are deterministically generated respectively as

$$W1 = \mathbb{I}[U_{W1} < 0.50]$$
$$W2 = \mathbb{I}[U_{W2} < 0.50]$$

  where $W1 = 1$ or $W2 = 1$ if the value of U_{W1} or U_{W2}, respectively, is less than 0.50, and $W1 = 0$ or $W2 = 0$ otherwise.

Since U_{W1} and U_{W2} are each distributed as $Uniform(0, 1)$ and thus equally likely to be any number from 0 to 1, the probability that $W1 = 1$ and the probability that $W2 = 1$ are each 0.50, i.e.,

$$P_0(W1 = 1) = P_0(W2 = 1) = 0.50.$$

Since 0 is the only other possible value for each of $W1$ and $W2$, and the total probability must sum to 1, the probability that $W1 = 0$ and the probability that $W2 = 0$ are also each 0.50:

$$P_0(W1 = 0) = 1 - P_0(W1 = 1) = 1 - 0.50 = 0.50$$
$$P_0(W2 = 0) = 1 - P_0(W2 = 1) = 1 - 0.50 = 0.50$$

Moreover, since $W1$ and $W2$ are independent in the working structural causal model and this particular data generating process, we can obtain their joint distribution $P_0(W1 = w1, W2 = w2)$ via multiplication:

$$P_0(W1 = w1, W2 = w2) = P_0(W1 = w1) * P_0(W2 = w2)$$

Moreover, since $P_0(W1 = w1) = P_0(W2 = w2) = 0.50$ for every possible value of $w1$ and $w2$, $P_0(W1 = w1, W2 = w2) = P_0(W1 = w1) * P_0(W2 = w2) = 0.50 * 0.50 = 0.25$ for every one of the four possible combinations of $w1$ and $w2$.

Thus to calculate the statistical estimand $\Psi(\mathbb{P}_0)$, we can plug the expressions for the above expected values of Y, $\mathbb{E}_0(Y|A, W1, W2)$, and the joint probability of $W1 = w1$ and $W2 = w2$, $P_0(W1 = w1, W2 = w2)$, into the G-computation formula outlined above in numbered equation (1):

$$\begin{aligned}
\Psi(\mathbb{P}_0) &= \sum_{w1,w2} [\mathbb{E}_0(Y|A = 1, W1 = w1, W2 = w2) - \mathbb{E}_0(Y|A = 0, W1 = w1, W2 = w2)]\mathbb{P}_0(W1 = w1, W2 = w2) \\
&= \sum_{w1,w2} ([logit^{-1}(-0.75 + W1 - 2W2 + 2.5(A = 1) + (A = 1)W1) - logit^{-1}(-0.75 + W1 - 2W2 + 2.5(A = 0) + (A = 0)W1)] * 0.25 \\
&= [logit^{-1}(-0.75 + 1 - 2 * 1 + 2.5 * 1 + 1 * 1) - logit^{-1}(-0.75 + 1 - 2 * 1 + 2.5 * 0 + 0 * 1)] * 0.25 \\
&\quad + [logit^{-1}(-0.75 + 1 - 2 * 0 + 2.5 * 1 + 1 * 1) - logit^{-1}(-0.75 + 1 - 2 * 0 + 2.5 * 0 + 0 * 1)] * 0.25 \\
&\quad + [logit^{-1}(-0.75 + 0 - 2 * 1 + 2.5 * 1 + 1 * 0) - logit^{-1}(-0.75 + 0 - 2 * 1 + 2.5 * 0 + 0 * 0)] * 0.25 \\
&\quad + [logit^{-1}(-0.75 + 0 - 2 * 0 + 2.5 * 1 + 1 * 0) - logit^{-1}(-0.75 + 0 - 2 * 0 + 2.5 * 0 + 0 * 0)] * 0.25 \\
&= 0.506905
\end{aligned}$$

This says that the strata-specific (i.e., (health-care-access-and-conflict-history-specific) conditional probability of survival for those who receive the combination prevention package, averaged with respect to the distribution of the baseline covariates (health care access and confict history), is 0.5069 higher than that of those who do not receive the combination prevention package. Since this data generating process satisfies the positivity assumption and under this data generating process the set of covariates $(W1, W2)$ satisfies the backdoor criterion, for this data generating process, $\Psi(\mathbb{P}_0)$ identifies the average treatment effect $\Psi^F(\mathbb{P}_{U,X})$, the difference between the counterfactual probability of survival if all children receive the combination prevention package and the counterfactual probabiilty of survival if all children do not receive the package.

3 Translate this data generating process into simulations

```
library(tidyverse)
```

3.1 First set the seed to 252.

```
set.seed(252)
```

3.2 Set the number of draws $n = 100,000$.

```
n = 100000
```

3.3 Sample n independent and identically distributed (i.i.d.) observations of random variable $O = (W1, W2, A, Y) \sim \mathbb{P}_0$.

```
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)

W1 <- as.numeric(U_W1 < 0.5)
W2 <- as.numeric(U_W2 < 0.5)
A <- as.numeric(U_A < plogis(-0.5+W1-1.5*W2))
Y <- as.numeric(U_Y < plogis(-0.75+W1-2*W2+2.5*A+A*W1))

X <- tibble(W1, W2, A, Y)
```

3.4 *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 $\Psi^F(\mathbb{P}_{U,X})$.

```
Y_1 <- as.numeric(U_Y < plogis(-0.75+W1-2*W2+2.5*1+1*W1))

Y_0 <- as.numeric(U_Y < plogis(-0.75+W1-2*W2+2.5*0+0*W1))

Psi_F <- mean(Y_1) - mean(Y_0)

Psi_F

## [1] 0.50707
```

The above result of $\Psi^F(\mathbb{P}_{U,X}) = 0.5071$ shows that under this data generating process and given the distribution of baseline covariates $W1$ and $W2$ in this population, the difference between the counterfactual probability of survival if all children receive the combination prevention package and the counterfactual probabiilty of survival if all children do not receive the package, i.e., the average treatment effect, is 0.5071.

3.5 Evaluate the positivity assumption.

For the positivity assumption to hold, there must be a positive (but less than one) 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 ($W2$), i.e., all of the following must hold:

$$0 < \mathbb{P}_0(A = 1|W1 = 1, W2 = 1) < 1$$
$$0 < \mathbb{P}_0(A = 1|W1 = 1, W2 = 0) < 1$$
$$0 < \mathbb{P}_0(A = 1|W1 = 0, W2 = 1) < 1$$
$$0 < \mathbb{P}_0(A = 1|W1 = 0, W2 = 0) < 1$$

Using this simulated data, we can check the positivity assumption by checking whether the mean of A within each possible stratum of $(W1, W2)$ in the simulated data is between 0 and 1 exclusive. (Checking for whether $\mathbb{P}_0(A = 0|W1 = 1, W2 = 1) < 1$ is equivalent to checking whether $0 < \mathbb{P}_0(A = 0|W1 = 1, W2 = 1)$, as A can only take values of 0 or 1.):

```
mean_A_W1_1_W2_1 <- mean(A[W1 == 1 & W2 == 1])
mean_A_W1_1_W2_1

## [1] 0.271355

mean_A_W1_1_W2_0 <- mean(A[W1 == 1 & W2 == 0])
mean_A_W1_1_W2_0

## [1] 0.6221695

mean_A_W1_0_W2_1 <- mean(A[W1 == 0 & W2 == 1])
mean_A_W1_0_W2_1

## [1] 0.1190666

mean_A_W1_0_W2_0 <- mean(A[W1 == 0 & W2 == 0])
mean_A_W1_0_W2_0

## [1] 0.3756981
```

Thus we have

- For $W1 = 1, W2 = 1$:
  - $\mathbb{P}_0(A = 1|W1 = 1, W2 = 1) = 0.2714$
  - $\mathbb{P}_0(A = 0|W1 = 1, W2 = 1) = 1 - \mathbb{P}_0(A = 1|W1 = 1, W2 = 1) = 0.7286$
- For $W1 = 1, W2 = 0$:
  - $\mathbb{P}_0(A = 1|W1 = 1, W2 = 0) = 0.6222$
  - $\mathbb{P}_0(A = 0|W1 = 1, W2 = 0) = 1 - \mathbb{P}_0(A = 1|W1 = 1, W2 = 0) = 0.3778$
- For $W1 = 0, W2 = 1$:
  - $\mathbb{P}_0(A = 1|W1 = 0, W2 = 1) = 0.1191$
  - $\mathbb{P}_0(A = 0|W1 = 0, W2 = 1) = 1 - \mathbb{P}_0(A = 1|W1 = 0, W2 = 1) = 0.8809$
- For $W1 = 0, W2 = 0$:
  - $\mathbb{P}_0(A = 1|W1 = 0, W2 = 0) = 0.3757$
  - $\mathbb{P}_0(A = 0|W1 = 0, W2 = 0) = 1 - \mathbb{P}_0(A = 1|W1 = 0, W2 = 0) = 0.6243$

Since the simulated mean probability of $A = 0$ and the simulated mean probability of $A = 1$ are greater than 0 for all strata (possible combinations of values of $W1$ and $W2$), the positivity assumption is met.

3.6 Evaluate the statistical estimand $\Psi(\mathbb{P}_0)$ and assign the value ψ_0 to `Psi.P0`.

```
mean_Y_A_1_W1_1_W2_1 <- mean(Y[A == 1 & W1 == 1 & W2 == 1])

mean_Y_A_0_W1_1_W2_1 <- mean(Y[A == 0 & W1 == 1 & W2 == 1])

P_W1_1_W2_1 <- length(Y[W1 == 1 & W2 == 1])/n


mean_Y_A_1_W1_1_W2_0 <- mean(Y[A == 1 & W1 == 1 & W2 == 0])

mean_Y_A_0_W1_1_W2_0 <- mean(Y[A == 0 & W1 == 1 & W2 == 0])

P_W1_1_W2_0 <- length(Y[W1 == 1 & W2 == 0])/n


mean_Y_A_1_W1_0_W2_1 <- mean(Y[A == 1 & W1 == 0 & W2 == 1])

mean_Y_A_0_W1_0_W2_1 <- mean(Y[A == 0 & W1 == 0 & W2 == 1])

P_W1_0_W2_1 <- length(Y[W1 == 0 & W2 == 1])/n


mean_Y_A_1_W1_0_W2_0 <- mean(Y[A == 1 & W1 == 0 & W2 == 0])

mean_Y_A_0_W1_0_W2_0 <- mean(Y[A == 0 & W1 == 0 & W2 == 0])

P_W1_0_W2_0 <- length(Y[W1 == 0 & W2 == 0])/n


# underscore instead of period because periods are of the devil

Psi_P0 <-
  (mean_Y_A_1_W1_1_W2_1 - mean_Y_A_0_W1_1_W2_1)*P_W1_1_W2_1 +
  (mean_Y_A_1_W1_1_W2_0 - mean_Y_A_0_W1_1_W2_0)*P_W1_1_W2_0 +
  (mean_Y_A_1_W1_0_W2_1 - mean_Y_A_0_W1_0_W2_1)*P_W1_0_W2_1 +
  (mean_Y_A_1_W1_0_W2_0 - mean_Y_A_0_W1_0_W2_0)*P_W1_0_W2_0

Psi_P0

## [1] 0.5041414
```

3.7 Interpret $\Psi(\mathbb{P}_0)$.

This says that the strata-specific (i.e., (health-care-access-and-conflict-history-specific) conditional probability of survival for those who receive the combination prevention package, averaged with respect to the distribution of the baseline covariates (health care access and confict history), is 0.5041 higher than that of those who do not receive the combination prevention package. Since this data generating process satisfies the positivity assumption and under this data generating process the set of covariates $(W1, W2)$ satisfies the backdoor criterion, if this is the real data generating process, $\Psi(\mathbb{P}_0)$ identifies the average treatment effect $\Psi^F(\mathbb{P}_{U,X})$, the difference between the counterfactual probability of survival if all children receive the combination prevention package and the counterfactual probabiilty of survival if all children do not receive the package.

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

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

```
R = 500

n = 200
```

4.2 Create a $R = 500$ by 4 matrix `estimates` to hold the resulting estimates obtained at each iteration.

```
estimates <- matrix(NA, nrow = 500, ncol = 4)
```

4.3 Inside a `for` loop from $r = 1$ to $r = R = 500$, 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 dataset `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 `control`.

d. Estimator 1: Use the `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:

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

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

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

$$\bar{Q}_0^2(A, W) = logit^{-1}(\beta_0 + \beta_1 A + \beta_2 W1)$$

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

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

$$\bar{Q}_0^3(A, W) = logit^{-1}(\beta_0 + \beta_1 A + \beta_2 W2)$$

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

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

$$\bar{Q}_0^4(A, W) = logit^{-1}(\beta_0 + \beta_1 A + \beta_2 W1 + \beta_3 W2 + \beta_4 A * W1 + \beta_5 A * W2)$$

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=control` and `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 intervention $\bar{Q}_n(0, W_i)$. Be sure to specify the arguments `newdata=control` and `type='response'`.

j. For *each* estimator of $\bar{Q}_0(A, W)$, estimate $\Psi(\mathbb{P}_0)$ 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:

$$\hat{\Psi}() = \frac{1}{n} \sum i = 1n[\bar{Q}_n(1, W_i) - \bar{Q}_n(0, W_i)]$$

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

```r
for(i in 1:R){

  # sample n i.i.d. observations of O = (W1, W2, A, Y)
  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)

  W1 <- as.numeric(U_W1 < 0.5)
  W2 <- as.numeric(U_W2 < 0.5)
  A <- as.numeric(U_A < plogis(-0.5+W1-1.5*W2))
  Y <- as.numeric(U_Y < plogis(-0.75+W1-2*W2+2.5*A+A*W1))

  # create data frame obs of the resulting observed data
  obs <- data.frame(W1, W2, A, Y)

  # copy the data set obs into two new data frames txt and control
  txt <- control <- obs

  # set A = 1 for all units in txt
  txt <- txt %>% mutate(A = 1)

  # set A = 0 for all units in control
  control <- control %>% mutate(A = 0)

  # estimator one (use glm to estimate conditional survival probability)
  estimator_one <- glm(Y ~ A, family = 'binomial', data = obs)

  # estimator two
  estimator_two <- glm(Y ~ A + W1, family = 'binomial', data = obs)

  # estimator three
  estimator_three <- glm(Y ~ A + W2, family = 'binomial', data = obs)

  # estimator four
  estimator_four <- glm(Y ~ A + W1 + W2 + A*W1 + A*W2,
                        family = 'binomial',
                        data = obs)
```

```
  # for each estimator predict expected mean outcome under the intervention
  predict_one_txt <- predict(estimator_one,
                             newdata = txt,
                             type = 'response')
  predict_two_txt <- predict(estimator_two,
                             newdata = txt,
                             type = 'response')
  predict_three_txt <- predict(estimator_three,
                               newdata = txt,
                               type = 'response')
  predict_four_txt <- predict(estimator_four,
                              newdata = txt,
                              type = 'response')

  # for each estimator predict expected mean outcome under the control
  predict_one_control <- predict(estimator_one,
                                 newdata = control,
                                 type = 'response')
  predict_two_control <- predict(estimator_two,
                                 newdata = control,
                                 type = 'response')
  predict_three_control <- predict(estimator_three,
                                   newdata = control,
                                   type = 'response')
  predict_four_control <- predict(estimator_four,
                                  newdata = control,
                                  type = 'response')

  # estimate psi_hat for each
  psi_hat_one <- mean(predict_one_txt - predict_one_control)
  psi_hat_two <- mean(predict_two_txt - predict_two_control)
  psi_hat_three <- mean(predict_three_txt - predict_three_control)
  psi_hat_four <- mean(predict_four_txt - predict_four_control)

  # assign the resulting values as a row in matrix estimates
  estimates[i,] <- c(psi_hat_one,
                     psi_hat_two,
                     psi_hat_three,
                     psi_hat_four)
}
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

5 Performance of the estimators

5.1 What is the average value of each estimator of $\Psi(\mathbb{P}_0)$ across $R = 500$ simulations?

```
mean_estimator_one <- mean(estimates[,1])
mean_estimator_one

## [1] 0.6505123

mean_estimator_two <- mean(estimates[,2])
mean_estimator_two

## [1] 0.6228431

mean_estimator_three <- mean(estimates[,3])
mean_estimator_three

## [1] 0.5653621

mean_estimator_four <- mean(estimates[,4])
mean_estimator_four

## [1] 0.5060037
```

- The average value of estimator one is $\mathbb{E}_0(\hat{\Psi}(\mathbb{P}_n^1)) = 0.65051$;
- The average value of estimator two is $\mathbb{E}_0(\hat{\Psi}(\mathbb{P}_n^2)) = 0.62284$;
- The average value of estimator three is $\mathbb{E}_0(\hat{\Psi}(\mathbb{P}_n^3)) = 0.56536$; and
- The average value of estimator four is $\mathbb{E}_0(\hat{\Psi}(\mathbb{P}_n^4)) = 0.506$.

5.2 Estimate the bias of each estimator.

```
bias_estimator_one <- mean(estimates[,1] - Psi_P0)
bias_estimator_one

## [1] 0.146371

bias_estimator_two <- mean(estimates[,2] - Psi_P0)
bias_estimator_two

## [1] 0.1187018

bias_estimator_three <- mean(estimates[,3] - Psi_P0)
bias_estimator_three

## [1] 0.06122073

bias_estimator_four <- mean(estimates[,4] - Psi_P0)
bias_estimator_four

## [1] 0.001862327
```

The bias, or mean difference between the point estimate $\hat{\Psi}(\mathbb{P}_n)$ and the true value of the statistical estimand $\Psi(\mathbb{P}_0)$, $\mathbb{E}_0[\hat{\Psi}(\mathbb{P}_n) - \Psi(\mathbb{P}_0)]$, for each of the estimators is

- $Bias(\hat{\Psi}(\mathbb{P}_n^1)) = \mathbb{E}_0[\hat{\Psi}(\mathbb{P}_n^1) - \Psi(\mathbb{P}_0)] = 0.14637$
- $Bias(\hat{\Psi}(\mathbb{P}_n^2)) = \mathbb{E}_0[\hat{\Psi}(\mathbb{P}_n^2) - \Psi(\mathbb{P}_0)] = 0.1187$

- $Bias(\hat{\Psi}(\mathbb{P}_n^3)) = \mathbb{E}_0[\hat{\Psi}(\mathbb{P}_n^3) - \Psi(\mathbb{P}_0)] = 0.06122$
- $Bias(\hat{\Psi}(\mathbb{P}_n^4)) = \mathbb{E}_0[\hat{\Psi}(\mathbb{P}_n^4) - \Psi(\mathbb{P}_0)] = 0.00186$

5.3 Estimate the variance of each estimator.

```
var_estimator_one <- var(estimates[,1])
var_estimator_one

## [1] 0.003184073

var_estimator_two <- var(estimates[,2])
var_estimator_two

## [1] 0.003727014

var_estimator_three <- var(estimates[,3])
var_estimator_three

## [1] 0.004709279

var_estimator_four <- var(estimates[,4])
var_estimator_four

## [1] 0.006161725
```

The variance of an etsimator is given by the mean value of the squares of the differences between the point estimates and their mean:

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

The variance for each of the estimators is:

- $Variance(\hat{\Psi}(\mathbb{P}_n^1)) = 0.00318$
- $Variance(\hat{\Psi}(\mathbb{P}_n^2)) = 0.00373$
- $Variance(\hat{\Psi}(\mathbb{P}_n^3)) = 0.00471$
- $Variance(\hat{\Psi}(\mathbb{P}_n^4)) = 0.00616$

5.4 Estimate the mean squared error (MSE) of each estimator.

```
# calculate mse for all four estimators
mse_estimator_one <- mean((estimates[,1] - Psi_P0)^2)
mse_estimator_one

## [1] 0.02460217

mse_estimator_two <- mean((estimates[,2] - Psi_P0)^2)
mse_estimator_two

## [1] 0.01780967
```

```r
mse_estimator_three <- mean((estimates[,3] - Psi_P0)^2)
mse_estimator_three

## [1] 0.008447838

mse_estimator_four <- mean((estimates[,4] - Psi_P0)^2)
mse_estimator_four

## [1] 0.00615287

# calculate mse equivalent bias^2 + var
mse_alternate_estimator_one <- bias_estimator_one^2 + var_estimator_one
mse_alternate_estimator_one

## [1] 0.02460853

mse_alternate_estimator_two <- bias_estimator_two^2 + var_estimator_two
mse_alternate_estimator_two

## [1] 0.01781712

mse_alternate_estimator_three <- bias_estimator_three^2 + var_estimator_three
mse_alternate_estimator_three

## [1] 0.008457257

mse_alternate_estimator_four <- bias_estimator_four^2 + var_estimator_four
mse_alternate_estimator_four

## [1] 0.006165194
```

The mean squared error (MSE) of an estimator is the mean of the squares of the differences between each point estimate $\hat{\Psi}(\mathbb{P}_n)$ and the true value of the statistical estimand $\Psi(\mathbb{P}_0)$, or, equivalently, the sum of the variance and the squared bias:

$$MSE(\hat{\Psi}(\mathbb{P}_n)) = \mathbb{E}_0((\hat{\Psi}(\mathbb{P}_n) - \Psi(\mathbb{P}_0)])^2)$$

The MSE for each of the estimators is:

- $MSE(\hat{\Psi}(\mathbb{P}_n^1)) = 0.0246$
- $MSE(\hat{\Psi}(\mathbb{P}_n^2)) = 0.0178$
- $MSE(\hat{\Psi}(\mathbb{P}_n^3)) = 0.0084$
- $MSE(\hat{\Psi}(\mathbb{P}_n^4)) = 0.0062$

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

The fourth estimator, which estimated $\bar{Q}_0^4(A, W)$ via the logistic regression model

$$\bar{Q}_0^4(A, W) = logit^{-1}(\beta_0 + \beta_1 A + \beta_2 W1 + \beta_3 W2 + \beta_4 A * W1 + \beta_5 A * W2),$$

performed the best (i.e., had the smallest mean squared error) of the four estimators. I am not surprised because I know the particular data generating process, by which the expected value of Y given the covariates A, $W1$, and $W2$ is

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

Thus I know the value of Y is generated by a logistic function of a constant intercept term, $W1$, $W2$, A, and the product of A and $W1$. Although all four estimators are logistic models, the first estimator only includes terms for the intercept and A, the second only terms for the intercept, A, and $W1$, the third only terms for the intercept, A, and $W2$. Only the fourth model includes both $W1$ and $W2$, and only the fourth model includes a product (interaction) term for A and $W1$. Thus only the fourth model is correctly parametrically specified and can capture the entire true data generation process. (It also contains a superfluous interaction term between A and $W2$, but the β coefficient for that term can be 0 to recover the correctly specified model for the true data generating process.)