R Homework Two

Katherine Wolf
Introduction to Causal Inference (PH252D)
March 10, 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.

2.2 Evaluate the statistical estimand $\Psi(\mathbb{P}_O)$ in closed form for this data generating process.

3 Translate this data generating process into simulations

3.1 First set the seed to 252.

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

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

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

3.5 Evaluate the positivity assumption.

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

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

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.

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

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 `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 Use `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 `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:

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

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

g. Estimator 4: Use `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:

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

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`.

5 Performance of the estimators

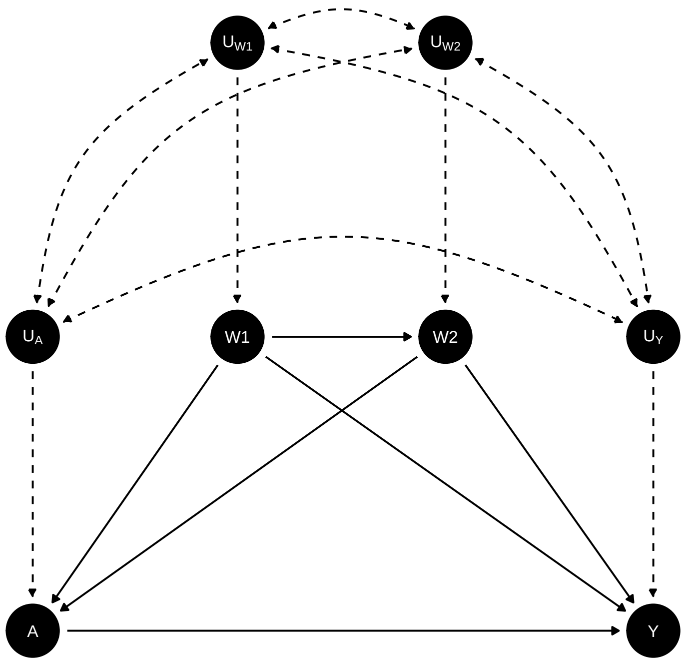
5.1 What is the average value of each estimator of $\Psi(\mathbb{P}_0)$ across $R = 500$ simulations?

5.2 Estimate the bias of each estimator.

5.3 Estimate the variance of each estimator.

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

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


U$_{W1}$
U$_{W2}$
U$_A$
W1
W2
U$_Y$
A
Y