STAT GR5293-005: Design and Analysis of Online Experiments Homework 4

Spring 2023

Due Apr 28 midnight (ET)

In this exercise, you will implement various estimators under the selection on observables (SOO) design using a simulated dataset.

The data generating process is described below:

- There are 3 control variables following uniform distributions
  - $x_1 \sim uniform(0,1)$, $x_2 \sim uniform(0,1)$, $x_3 \sim uniform(0,1)$
- The treatment indicator for each unit *i* is generated based on the following model
  $D_i = 1 \; if \; x_{i1} + x_{i2} + x_{i3} + e_i > 2; \; 0 \; otherwise$
  where $e_i \sim N(0,1)$
- The observed outcome is generated by the following model
  $Y_i = D_i + x_{i1} + x_{i2} + x_{i3} + \epsilon_i$ where $\epsilon_i = 2 * \eta_i$ and $\eta_i \sim N(0,1)$

From this data generating process, we know the true effect of D on Y is 1. Now, you need to draw 200 bootstrapped samples (with replacement) of size 500 from this simulated dataset (this will minimic the process of random sampling from a population) and estimate the average treatment effect (ATE) using the following approaches for each bootstrapped sample. Then you will get a distribution of ATE estimates per estimator. The key assumption of selection on observables (SOO) design is that we can control for all the factors affecting the treatment assignment. Hence, in this exercise, we will compare the effect estimates (in terms of mean and variance) when this assumption holds vs when this assumption breaks.

Estimator 0: Unadjusted means

$Y_i = \beta_0 + \beta_1 D_i + u_i$ and $\widehat{\beta_1}$ estimates the treatment effect (which is equivalent to comparing the unadjusted sample means between treatment and control. Note u_i is the error term.

Estimator 1.1: Regression adjustment with full control variables

$Y_i = \beta_0 + \beta_1 D_i + \beta_2 x_{i1} + \beta_3 x_{i2} + \beta_4 x_{i3} + u_i$ and $\widehat{\beta_1}$ estimates the treatment effect. Note this corresponds to a correctly specified model.

Estimator 1.2: Regression adjustment with partial control variables

$Y_i = \beta_0 + \beta_1 D_i + \beta_2 x_{i1} + u_i$ and $\widehat{\beta_1}$ estimates the treatment effect. This setup mimics the model misspecification.

Estimator 2.1: Conditioning on propensity score with full control variables

Include the estimated propensity score $\widehat{p_i}$ as a control variable and $\widehat{\beta_1}$ estimates the treatment effect:

$$Y_i = \beta_0 + \beta_1 D_i + \beta_2 \widehat{p_i} + u_i$$

Use logistics regression to model propensity scores, and include the all three variables x_1, x_2, x_3 as control variables: $logit(D) \sim x_1 + x_2 + x_3$ and generated the predicted propensity score $\widehat{p_i}$ for each observation *i*.

Estimator 2.2: Conditioning on propensity score with partial control variables

Include the estimated propensity score $\widehat{p_i}$ as a control variable and $\widehat{\beta_1}$ estimates the treatment effect:

$$Y_i = \beta_0 + \beta_1 D_i + \beta_2 \widehat{p_i} + u_i$$

Use logistics regression to model propensity scores, and include only x_1 as control variables: $logit(D) \sim x_1$ and generated the predicted propensity score $\widehat{p_i}$ for each observation *i* . This setup mimics the model misspecification.

Estimator 3.1: Re-weighting based on propensity score with full control variables

The estimated treatment effect is:

$\widehat{\tau} = \frac{1}{N}\left(\sum_{i=1}^{N} \frac{D_i Y_i}{\widehat{p_i}} - \sum_{i=1}^{N} \frac{(1-D_i)Y_i}{1-\widehat{p_i}}\right)$ where N is the total number of observations and $\widehat{p_i}$ is modeled using full control variables as described in estimator 2.1.

Estimator 3.2: Re-weighting based on propensity score with partial control variables

The estimated treatment effect is:

$\widehat{\tau} = \frac{1}{N}\left(\sum_{i=1}^{N} \frac{D_i Y_i}{\widehat{p_i}} - \sum_{i=1}^{N} \frac{(1-D_i)Y_i}{1-\widehat{p_i}}\right)$ where N is the total number of observations and $\widehat{p_i}$ is modeled using partial control variables as described in estimator 2.2.

Estimator 4.1: Blocking based on propensity score with full control variables

We divide the data into K groups according to $\widehat{p_i}$ and estimate a separate treatment effect $\widehat{\tau_k}$ for each group k. Then combine the treatment effects into a single ATE as

$\hat{\tau} = \sum_{k=1}^{K} \hat{\tau}_k \frac{N_k}{N}$ where N_k is the total number of observations from both treatment and control that fall into the k group and N is the total number of observations. $\hat{p}_i$ is modeled using full control variables as described in estimator 2.1.

You can use 0.05 increment to define a block. Another restriction is on the overlap of $\hat{p}_i$ between treatment and control. Find the min and max of estimated propensity scores in both treatment and control groups, call them $\hat{p}_{max\ trt}$, $\hat{p}_{min\ trt}$, $\hat{p}_{max\ ctl}$, $\hat{p}_{min\ ctl}$. Make sure you drop observations with estimated propensity scores above min($\hat{p}_{max\ trt}$, $\hat{p}_{max\ ctl}$). Similarly, drop observations with the estimated propensity score below max($\hat{p}_{min\ trt}$, $\hat{p}_{min\ ctl}$).

Estimator 4.2: Blocking based on propensity score with partial control variables

We divide the data into K groups according to $\hat{p}_i$ and estimate a separate treatment effect $\hat{\tau}_k$ for each group k. Then combine the treatment effects into a single ATE as

$\hat{\tau} = \sum_{k=1}^{K} \hat{\tau}_k \frac{N_k}{N}$ where N_k is the total number of observations from both treatment and control that fall into the k group. $\hat{p}_i$ is modeled using partial control variables as described in estimator 2.2. Make sure you implement the same overlap restriction as stated in estimator 4.1.

Estimator 5.1: Doubly robust estimator with full control variables

$Y_i = \beta_0 + \beta_1 D_i + \beta_2 x_{i1} + \beta_3 x_{i2} + \beta_4 x_{i3} + u_i$ where the regression weights are defined as $w_i = \sqrt{\frac{D_i}{\hat{p}_i} + \frac{1-D_i}{1-\hat{p}_i}}$ and $\hat{p}_i$ is modeled using full control variables as described in estimator 2.1.

Estimator 5.2: Doubly robust estimator with partial control variables

$Y_i = \beta_0 + \beta_1 D_i + \beta_2 x_{i1} + \beta_3 x_{i2} + \beta_4 x_{i3} + u_i$ where the regression weights are defined as $w_i = \sqrt{\frac{D_i}{\hat{p}_i} + \frac{1-D_i}{1-\hat{p}_i}}$ and $\hat{p}_i$ is modeled using partial control variables as described in estimator 2.2.

Problem 1: Create a table using the following template for all the estimators described above (the table below is just an example)

	Mean of ATE	95% CI of ATE

Estimator 0		
Estimator 1.1		

Problem 2: Let's look at cases when the model is being correctly specified. Plot the ATE estimates distribution for Estimator 0, 1.1, 2.1, 3.1, 4.1, 5.1 in one graph and report your findings.

- Are there any differences among those ATE estimates in terms of mean and variance of the estimates?
- What is the comparison with the true ATE?

Problem 3: Let's look at cases when the model is being incorrectly specified. Plot the ATE estimates distribution for Estimator 0, 1.2, 2.2, 3.2, 4.2, 5.2 in one graph and report your findings.

- Are there any differences among those ATE estimates in terms of mean and variance of the estimates?
- How do they compare to the true ATE?
- Note in Estimator 5.2, the propensity score model is incorrectly specified, but the conditional expectation function of y is correctly specified. Are you seeing the benefit of using a double robust estimator?

Problem 4: Compare the results between problem 2 and problem 3. Comments on the impact of the violation of SOO design assumptions.

Problem 5: Let's assess the impact of overlap. Zoom into estimator 4.1 and 4.2. Let's remove the overlapping restriction (e.g. don't drop any observations) and re-estimate ATE.

- What are the differences between these estimates vs the previous estimates in terms of mean and variance? Explain why there exists such differences.
- What is the impact of ignoring overlapping restrictions comparing correctly specified models vs misspecified models?