ATE in randomized controlled trials and observational studies

In this Lab you will explore the main differences in the ATE's estimation from RCT and observational data. You will study the properties of the IPW and parametric g-formulaÄôs estimators trough extensive simulations on synthetic data simulated as in (Lunceford & Davidian 2004).

Synthetic data - (Lunceford & Davidian 2004)

We will perform the same simulations proposed in Lunceford & Davidian 2004.

Generative model

The response variable Y is generated according the following equation:

$$Y = \nu_0 + \nu_1 X_1 + \nu_2 X_2 + \nu_3 X_3 + \nu_4 A + \xi_1 V_1 + \xi_2 V_2 + \xi_3 V_3 + \varepsilon, \varepsilon \sim \mathcal{N}(0,1),$$

where $\nu = (\nu_0, \nu_1, \nu_2, \nu_3, \nu_4)^T = (0, -1, 1, -1, 2), \xi = (\xi_1, \xi_2, \xi_3) = (-1, 1, 1)$. The covariates are distributed as $X_3 \sim Bernoulli(0.2)$, and conditionnaly on X_3

- If $X_3 = 0$, $V_3 \sim Bernoulli(0.25)$ and $(X_1, V_1, X_2, V_2)^T \sim \mathcal{N}(\tau_0, \Sigma)$
- If $X_3 = 1$, $V_3 \sim Bernoulli(0.75)$ and $(X_1, V_1, X_2, V_2)^T \sim \mathcal{N}(\tau_1, \Sigma)$

with

$$\tau_0 = \begin{pmatrix} -1 \\ -1 \\ 1 \\ 1 \end{pmatrix}, \tau_1 = \begin{pmatrix} 1 \\ 1 \\ -1 \\ -1 \end{pmatrix}, \Sigma = \begin{pmatrix} 1 & 0.5 & -0.5 & -0.5 \\ 0.5 & 1 & -0.5 & -0.5 \\ -0.5 & -0.5 & 1 & 0.5 \\ -0.5 & -0.5 & 0.5 & 1 \end{pmatrix}$$

and the treatment A is generated as a Bernoulli of the propensity score

$$e(X, \beta) = \frac{1}{1 + \exp(-\beta_0 - \beta_1 X_1 - \beta_2 X_2 - \beta_3 X_3)}$$

with $\beta = (\beta_0, \beta_1, \beta_2, \beta_3)^T = (0, 0.6, -0.6, 0.6)$.

The data generating process is implemented in the Python notebook given in Moodle.

Question 1

What is the true ATE?

Question 2

Represent the distribution of the covariates X_1 , X_2 , X_3 with respect to the treatment? Is the data set balanced?

Parametric estimation

You will implement (1) the parametric Inverse Probability Weighting estimator, as well as (2) the OLS estimator. You will study how different sets of covariates as well as model misspecification affect the ATE estimation.

Question 3

(3.1) Estimate the ATE using the difference in means estimator. What do you observe?
(3.2) Are there any confounding variables in the proposed model? To answer the question you can draw a so-called causal model as a DAG.

Question 4

(4.1) Estimate the ATE using the conditional means OLS estimator (parametric g formula

estimator). First, use only the set $\{X_1, X_2, X_3\}$ for adjustment.
(4.2) Fit the model using only X_1 and X_3 . What do you observe?

Question 5

(5.1) Implement the parametric IPW estimator. Fit the propensity score using $\{X_1, X_2, X_3\}$.
(5.2) Fit the model using only X_1 and X_3. What do you observe?

Question 6

Represent the distribution of the propensity score weights. Is overlap verified?
An important step to check that the estimation of the propensity scores (also called the weights) helps the remove confounding is to assess whether or not the weights allows to achieve a better balance between treated and control units. Achieving balance justifies ignorability on the observed covariates, allowing for the potential for a valid causal inference after effect estimation. Some authors also advocate that reporting a good balance belongs to good practice to ensure that confounding bias can be removed. Here we suggest to use the Python package `psmpy` to assess the balance.

Question 7

(7.1) Illustrate the properties of the different estimators by performing a simulation. You can add the g-estimator with all the variables X and V , what do you observe?
Tip: repeat 100 simulatioons, with 5000 datapoints per simulation
(7.2) Answer Question (7.1) using different number of datapoints per simulation (for instance 100, 1000, 10000). Comment.

Nonparametric

Question 8

Repeat the analysis (Question 7) but using non-parametric estimation strategies, for instance using Random Forest to estimate the nuisance components. Comment the results.