Treatment effect estimation

Potential outcome framework (Neyman, 1923, Rubin, 1974)

- n iid samples $(X_i, T_i, Y_i(0), Y_i(1)) \in \mathbb{R}^d \times \{0,1\} \times \mathbb{R} \times \mathbb{R}$
- Treatment: intervention, "can be manipulated"
- Note $Y_i = Y_i(T_i)$, the observed data is: (Y_i, X_i, T_i)
- One has $Y_i = T_i Y_i(1) + (1 - T_i) Y_i(0)$
- PO vs counterfactuals

X_1	X_2	X_3	T	Y	Y(0)	Y(1)
5	1	F	1	10	?	10
-1	2	M	1	5	?	5
⋮	⋮	⋮	⋮	⋮	⋮	⋮
…	…	…	0	6	6	?
…	…	…	1	8	?	8
6	4	M	0	4	4	?

Potential outcome framework (Neyman, 1923, Rubin, 1974)

- Individual causal effect of the treatment: $\Delta_i := Y_i(1) - Y_i(0)$
- Missing problem: Δ_i never observed (only observe one outcome/indiv)
- Average treatment effect (ATE)

$$\tau = \mathbb{E}[\Delta_i] = \mathbb{E}[Y_i(1) - Y_i(0)]$$

- How can we calculate the ATE ?
- One convenient framework : Random treatment assignment (RCT)

$$T_i \perp\!\!\!\perp (X_i, Y_i(0), Y_i(1))$$

Randomized Controlled Trial (A/B testing)

Identifiability assumptions

- $Y_i = T_i Y_i(1) + (1 - T_i) Y_i(0)$ (consistency)
- $T_i \perp\!\!\!\perp (Y_i(0), Y_i(1), X_i)$ (random treatment assignment)

One can check that

$$\begin{aligned} \tau &= \mathbb{E}[\Delta_i] = \mathbb{E}[Y_i(1)] - \mathbb{E}[Y_i(0)] \\ &= \mathbb{E}[Yi(1) | T_i = 1] - \mathbb{E}[Y_i(0) | T_i = 0] \text{ (RCT)} \\ &= \mathbb{E}[Y_i | T_i = 1] - \mathbb{E}[Y_i | T_i = 0] \text{ (consistency)} \end{aligned}$$

Difference-in-means estimator

$$\widehat{\tau}_{DM} = \frac{1}{n_1} \sum_{T_i=1} Y_i - \frac{1}{n_0} \sum_{T_i=0} Y_i$$

$\widehat{\tau}_{DM}$ unbiased and $\sqrt{n}$ consistent (CLT satisfied)

Randomized Controlled Trial (A/B testing)

Identifiability assumptions

- $Y_i = T_i Y_i(1) + (1 - T_i) Y_i(0)$ (consistency)
- $T_i \perp\!\!\!\perp (Y_i(0), Y_i(1), X_i)$ (random treatment assignment)

One defines

$$\widehat{\tau}_{DM} = \frac{1}{n_1} \sum_{T=1} Y_i - \frac{1}{n_0} \sum_{T=0} Y_i$$

T	Y	Y(0)	Y(1)
1	10	?	10
0	6	6	?
1	8	?	8
0	4	4	?

ATE=mean(orange)-mean(green)

Beyond RCT

A randomized experiment is an assignment mechanism such that:

- The assignment mechanism is ignorable: the assignment mechanism does not depend on the counterfactual outcomes, that is,

$$\mathbb{P}[T = 1|X, Y(0), Y(1)] = \mathbb{P}[T = 1|X, Y_{obs}]$$

- The assignment mechanism is probabilistic: the probability of treatment assignment to a unit satisfies

$$0 < \mathbb{P}[T = 1|X, Y(0), Y(1)] < 1$$

- The assignment mechanism is a known function of its arguments.
- A randomized controlled trial is a randomized experiment such that

$$T \perp\!\!\!\perp (X, Y(0), Y(1))$$

Beyond RCT

- An assignment mechanism corresponds to an observational study if it is an unknown function of its arguments.
- In practice, lack of a controlled design for a lot of experimental data in epidemiological studies, insurance claims, administrative data
- We need to go beyond RCT and to take into account the selection bias when collecting the data
- Need of an appropriate mathematical framework to deal with more complex situations

Beyond RCT[1]

A motivating example

- Beyond one RCT? Two RCTs!
- Supposed that we are interested in giving teenagers cash incentives to discourage them from smoking.
- Two populations are mixed: 5% of teenagers in Palo Alto, CA, 20% of teenagers in Geneva, Switzerland.
- Effect of the treatment, i.e. of the incentive

[1] Chap 1 of "Causal Inference" course of S. Wager

Beyond RCT[1]

Palo Alto	Non S.	Smoker	Geneva	Non S.	Smoker
Treat.	152	5	Treat.	581	350
Contr.	2362	122	Contr.	2278	1979

$$\widehat{\tau}_{PA} = 5/(152+5) - 122/(2362+122) = -0.02$$

$$\widehat{\tau}_{GVA} = 350/(350+581) - 1979/(2278+1979) = -0.09$$

The treatment seems to help!

[1] Chap 1 of "Causal Inference" course of S. Wager

Beyond RCT[1]

Have a look at data aggregated in a naive way: Simpson paradox

Palo Alto+Geneva	Non S.	Smoker
Treat.	733	401
Contr.	4640	2101

$$\widehat{\tau}_{naive} = 401/(733+401) - 2101/(2101+4640) = 0.04$$

The treatment seems to hurt!

[1] Chap 1 of "Causal Inference" course of S. Wager

Beyond RCT[1]

- After aggregating the data, no longer an RCT
- Genevans are both more likely to get treated, and also more likely to smoke
- In order to get a consistent estimate of the ATE, we need to estimate treatment effects in each city separately

[1] Chap 1 of "Causal Inference" course of S. Wager

Beyond RCT[1]

- Solve the Simpson paradox?
- Proposed estimator

$$\widehat{\tau}_{smart} = \frac{2641}{2641+5188}\widehat{\tau}_{PA} + \frac{5188}{2641+5188}\widehat{\tau}_{GVA} = -0.06$$

- We are now consistent with the results for each city!

[1] Chap 1 of "Causal Inference" course of S. Wager

Beyond RCT[1]

Key ideas

:

- Divide the population into relevant groups corresponding to a levl of the covariates X
- Define an estimator $\widehat{\tau}(x)$ of the groupwise ATE corrsponding to $X = x$

$$\tau(x) = \mathbb{E}[Y_i(1) - Y_i(0)|X = x]$$

- Combine these groupwise ATEs as follows

$$\widehat{\tau}_{agg} = \sum_{x} \frac{n_x}{n} \widehat{\tau}(x)$$

[1] Chap 1 of "Causal Inference" course of S. Wager

Beyond RCT[1]

- How can we generalize this approach?
- Consider an appropriate mathematical framework

[1] Chap 1 of "Causal Inference" course of S. Wager

Assumption for ATE identifiability in observational data

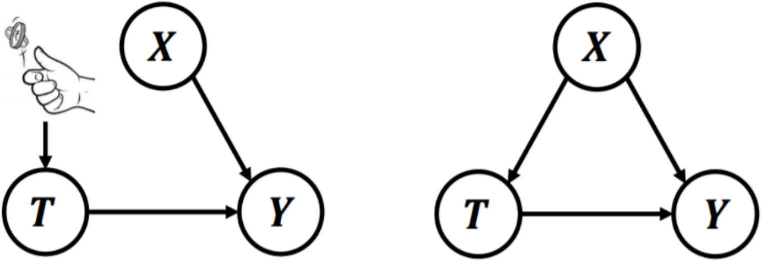
Unconfoundedness

$$T_i \perp\!\!\!\perp (Y_i(0), Y_i(1)) | X_i$$

- Treatment assignment T_i is random conditionally on covariates X_i
- Measure enough covariates to capture dependence between T_i and outcomes
- Generalize the “RCT in each group” assumption

Assumption for ATE identifiability in observational data

Interpretation in term of graphical model



Causal structure for RCT and observational studies [Li et al., 2020]

Assumption for ATE identifiability in observational data

Confounder

- A confounder is a third variable that is related to both the exposure of interest and the response.
- The effect of the treatment may be confounded by factors related to which group the subjects were assigned.

ATE not identifiable without unconfoundness assumption

- It is not a sample size problem, i.e., w/o it we cannot identify ATE even with infinite amount of data.
- Unobserved confounders makes it impossible to separate correlation and causality
- Assumption not testable from the data.

The propensity score

- The ucounfoundness assumption involves a covariate vector X_i that may be high dimensional
- The situation can be simplified using the so called propensity score
$$e(X) = \mathbb{P}[T_i = 1 | X_i = x]$$

We assume overlap, i.e. for some $\eta > 0$,

$$\eta < e(x) < 1 - \eta$$

The propensity score

Balancing property of the propensity score (Rubin et al. 1983)

The uncounfoundness assumption implies that

$$T_i \perp\!\!\!\perp (Y_i(0), Y_i(1)) | e(X_i)$$

Proof : See Chap 2 of Course "Causal Inference" of S. Wager

The propensity score

Consequences

- It suffices to control for $e(X)$ (rather than X), to remove biases associated with non-random treatment assignment.
- Matching procedure using a dimensionality reduction step?
- One can compare observations with same ps with different covariates.

Two methods to estimate the ATE

- Propensity stratification
- Inverse propensity weigthing

The propensity score

Algorithm of propensity stratification

Step 1 : preliminaries

- Step 1-a : obtain an estimate $\widehat{e}(x)$ of the propensity score
- Step 1-b : choose a number of strata J

Step 2: define strata

- Sort the observations according to their propensity scores

$$\widehat{e}(X_{i_1}) \leq \widehat{e}(X_{i_2}) \leq \cdots \leq \widehat{e}(X_{i_n}).$$

- Step 2-b : Split the sample into J evenly size strata using the sorted propensity score

The propensity score

Algorithm of propensity stratification

Step 3 : Estimate the average treatment

- Step 3-a : In each stratum $j = 1, \cdots, J$ compute the simple difference-in-means treatment effect estimator $\widehat{\tau}_j$ for the stratum
- Step 3-b : Define the aggregated estimator

$$\widehat{\tau}_{strat} = \frac{1}{J} \sum_j \widehat{\tau}_j$$

The propensity score

Algorithm of propensity stratification

Properties of $\widehat{\tau}_{strat}$

- Estimator consistent if $\widehat{e}(X)$ uniformly consistent for $e(x)$
- CLT under suitable condition for J

The propensity score

Inverse Propensity Weighting

Identifiability with Propensity Score

$$
\begin{aligned}
\tau &= \mathbb{E}[Y_i(1) - Y_i(0)] \\
&= \mathbb{E}[\mathbb{E}[Y_i(1)|X] - \mathbb{E}[Y_i(0)|X_i]] \\
&= \mathbb{E}\left[\frac{\mathbb{E}[T_i|X_i] \cdot \mathbb{E}[Y_i(1)|X]}{e(X_i)} - \frac{\mathbb{E}[1-T_i|X_i]_i(0)|X_i]}{1-e(X_i)}\right] \quad \text{def. of } e(x) \\
&= \mathbb{E}\left[\frac{\mathbb{E}[T_i \cdot Y_i(1)|X_i]}{e(X_i)} - \frac{\mathbb{E}[(1-T_i) \cdot Y_i(0)|X_i]}{1-e(X_i)}\right] \quad \text{uncounfoundness} \\
&= \mathbb{E}\left[\frac{T_i \cdot Y_i}{e(X_i)} - \frac{(1-T_i) \cdot Y_i}{1-e(X_i)}\right]
\end{aligned}
$$

For the last equality, we use

$$TY = T(TY(1) + (1-T)Y(0)) = T^2Y(1) + T(1-T)Y(0) = TY(1).$$

The propensity score

Inverse Propensity Weighting

IPW oracle estimator

One deduces from this heuristic analysis the definition of the following so called IPW oracle estimator

$$\widehat{\tau}^*_{IPW} = \frac{1}{n}\sum_i \left[\frac{T_i \cdot Y_i}{e(X_i)} - \frac{(1-T_i)\cdot Y_i}{1-e(X_i)}\right]$$

- Weighting subjects by the inverse probability of treatment received creates a synthetic sample in which treatment assignment is independent of covariates
- The process of weighting by the inverse probability of treatment allowed to adequately balance the major differences between the two groups

The propensity score

Inverse Propensity Weighting

Set $m(X) = \mathbb{E}[Y_i | X_i = x]$ and $\tau(X) = \mathbb{E}[Y_i(1) - Y_i(0) | X_i = x]$.
One has

Large sample properties

$$\sqrt{n}(\widehat{\tau}^*_{IPW} - \tau) \to \mathcal{N}(0, V^*_{IPW})$$

with

$$V^*_{IPW} = \mathbb{E}\left[\frac{m^2(X)}{e(X)(1-e(X))}\right] + var(\tau(X)) + \mathbb{E}\left[\frac{\sigma^2(X)}{e(X)(1-e(X))}\right]$$

The propensity score

Inverse Propensity Weighting

- Let $\widehat{e}$ a consistent estimator of e
- Set

$$\widehat{\tau}_{IPW} = \frac{1}{n}\sum_i \left[\frac{T_i \cdot Y_i}{\widehat{e}(X_i)} - \frac{(1 - T_i) \cdot Y_i}{1 - \widehat{e}(X_i)} \right]$$

Theorem: IPW consistency

Assume that

- (H1) $\sup |\widehat{e}(x) - e(x)| \overset{a.s.}{\to} 0$ when $n \to \infty$
- (H2) Y is square integrable

Then

$$\left|\widehat{\tau}_{IPW,n} - \tau\right| \overset{a.s.}{\to} 0 \text{ as } n \to \infty$$

The propensity score

Linear model in observational data

Given unconfoundedness

$$\begin{aligned}
\tau &= \mathbb{E}[\Delta_i] = \mathbb{E}[Y_i(1)] - \mathbb{E}[Y_i(0)] \\
&= \mathbb{E}[\mathbb{E}[\Delta_i|X_i]] \\
&= \mathbb{E}[\mathbb{E}[Y_i(1)|X_i]] - \mathbb{E}[\mathbb{E}[Y_i(0)|X_i]] \\
&= \mathbb{E}[\mu_1(X_i)] - \mathbb{E}[\mu_0(X_i)] \\
&= \mathbb{E}[\mathbb{E}[Y_i(1)|T_i = 1, X_i]] - \mathbb{E}[\mathbb{E}[Y_i(0)|T_i = 0, X_i]] \\
&= \mathbb{E}[\mathbb{E}[Y_i|T_i = 1, X_i]] - \mathbb{E}[\mathbb{E}[Y_i|T_i = 0, X_i]]
\end{aligned}$$

This suggest an estimator based on the difference between estimated conditional expectation

The propensity score

Linear model in observational data

Linearity of the responses $Y_i(0)$ and $Y_i(1)$ in the covariates

- $Y_i(t) = c(t) + X_i\beta(t) + \varepsilon_i(t),\ t \in \{0,1\}$
- $\mathbb{E}[\varepsilon_i(t)|X_i] = 0$ and $Var(\varepsilon_i(t)|X_i) = \sigma^2$.

OLS estimator

$$\begin{aligned} \widehat{\tau}_{OLS} &= \frac{1}{n}\sum_i (\widehat{\mu}_1(X_i) - \widehat{\mu}_0(X_i)) \\ &= \frac{1}{n}\sum_i [(\widehat{c}_1 + \widehat{\beta}_1 X_i) - (\widehat{c}_0 + \widehat{\beta}_0 X_i)] \end{aligned}$$

The propensity score

Linear model in observational data

One can add a lasso penalty in the estimation [1]

1. Run a LASSO of T on X . Select variables with non-zero coefficients at a selected (e.g. cross-validation).
2. Run a LASSO of Y on X on both the treated on control samples. Select variables with non-zero coefficients at a selected (may be different than first).
3. Run OLS of Y on T interacted with the union of selected variables. Conclude as in the regular OLS case.

The third step above is not as good at purely predicting Y as using only second set. But it is more accurate for the ATE. Result: under approximate sparsity of BOTH the propensity and outcome models, and constant treatment effects, estimated ATE is asymptotically normal and estimation is efficient

[1] Belloni, Chernozukov, and Hansen (2014)