Causal Inference

An introduction based on S. Wager's course on Causal Inference (OIT 661)

Imke Mayer

November 23, 2018

Group Meeting, CMAP

Outline

1. Treatment effect estimation in randomized experiments
2. Beyond a single randomized controlled trial
3. Inverse-propensity weighting
4. Double robustness property
5. Cross-fitting and machine learning for ATE estimation
6. Conclusion

Treatment effect estimation in randomized experiments

Definitions and notations

Given a treatment, define the **causal effect** via potential outcomes:

Causal effect

Binary treatment $w \in \{0, 1\}$ on *i-th* individual with potential outcomes $Y_i(1)$ and $Y_i(0)$.

Individual causal effect of the treatment:

$$\Delta_i = Y_i(1) - Y_i(0)$$

Definitions and notations

Given a treatment, define the **causal effect** via potential outcomes:

Causal effect

Binary treatment $w \in \{0, 1\}$ on *i-th* individual with potential outcomes $Y_i(1)$ and $Y_i(0)$.

Individual causal effect of the treatment:

$$\Delta_i = Y_i(1) - Y_i(0)$$

- Problem: Δ_i never observed.
- (Partial) Solution: randomized experiments to learn certain properties of Δ_i.
- **Average treatment effect** $\tau = \mathbb{E}[\Delta_i] = \mathbb{E}[Y_i(1) - Y_i(0)]$.

Average treatment effect (ATE)

Average treatment effect

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

Idea: estimate τ using large randomized experiments.

Assumptions:

Random variables (Y, W) having values in $\mathbb{R} \times \{0, 1\}$.

Observe n iid samples (Y_i, W_i) each satisfying:

- $Y_i = Y_i(W_i)$ (SUTVA)
- $W_i \perp\!\!\!\perp \{Y_i(0), Y_i(1)\}$ (random treatment assignment)

Average treatment effect (ATE)

Average treatment effect

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

Idea: estimate τ using large randomized experiments.

Assumptions:

Random variables (Y, W) having values in $\mathbb{R} \times \{0, 1\}$.

Observe n iid samples (Y_i, W_i) each satisfying:

- $Y_i = Y_i(W_i)$ (SUTVA)
- $W_i \perp\!\!\!\perp \{Y_i(0), Y_i(1)\}$ (random treatment assignment)

Difference-in-means estimator

$$\hat{\tau}_{DM} = \frac{1}{n_1} \sum_{W_1=1} Y_i - \frac{1}{n_0} \sum_{W_1=0} Y_i,$$

where $n_w = \#\{i : W_i = w\}$.

Average treatment effect estimation

Properties of $\hat{\tau}_{DM}$

- Using the previous assumptions (iid, SUTVA, random treatment assignment), we can prove that $\hat{\tau}_{DM}$ is unbiased and $\sqrt{n}$-consistent.

$$\sqrt{n}\left(\hat{\tau}_{DM} - \tau\right) \xrightarrow[n\to\infty]{d} \mathcal{N}(0, V_{DM}),$$

where $V_{DM} = \frac{Var(Y_i(0))}{\mathbb{P}(W_i=0)} + \frac{Var(Y_i(1))}{\mathbb{P}(W_i=1)}$

Average treatment effect estimation

Properties of $\hat{\tau}_{DM}$

- Using the previous assumptions (iid, SUTVA, random treatment assignment), we can prove that $\hat{\tau}_{DM}$ is unbiased and $\sqrt{n}$-consistent.

$$\sqrt{n}\,(\hat{\tau}_{DM} - \tau) \xrightarrow[n\to\infty]{d} \mathcal{N}(0, V_{DM}),$$

  where $V_{DM} = \frac{Var(Y_i(0))}{\mathbb{P}(W_i=0)} + \frac{Var(Y_i(1))}{\mathbb{P}(W_i=1)}$

- Using plug-in estimators we also get confidence intervals

$$\lim_{n\to\infty} \mathbb{P}\left(\tau \in \left(\hat{\tau}_{DM} \pm \Phi^{-1}(1-\alpha/2)\sqrt{\frac{\hat{V}_{DM}}{n}}\right)\right) = 1-\alpha,$$

  where Φ is the standard Gaussian cdf.

Average treatment effect estimation with Difference-of-Means

Difference-of-Means estimator

- conceptually simple estimator and simple to estimate,
- consistent estimator with asymptotically valid inference,
- but is it the optimal way to use the data for fixed finite n?

Average treatment effect estimation with Difference-of-Means

Difference-of-Means estimator

- conceptually simple estimator and simple to estimate,
- consistent estimator with asymptotically valid inference,
- but is it the optimal way to use the data for fixed finite n?

Average Treatment effect

τ is a **causal parameter**, i.e. property we wish to know about a population. It is not related to the study design or the estimation method.

Randomized trials in the linear model

Idea: assume linearity of the responses $Y_i(0)$ and $Y_i(1)$ in the covariates.

Assumptions

- n iid samples (X_i, Y_i, W_i),
- $Y_i(w) = c_{(w)} + X_i\beta_{(w)} + \varepsilon_i(w)$, $w \in \{0, 1\}$,
- $\mathbb{E}[\varepsilon_i(w)|X_i] = 0$ and $Var(\varepsilon_i(w)|X_i) = \sigma^2$.

and without loss of generality we additionally assume:

- $\mathbb{P}(W_i = 0) = \mathbb{P}(W_i = 1) = \frac{1}{2}$,
- $\mathbb{E}[X] = 0$.

Randomized trials in the linear model

OLS estimator

$$\begin{aligned}\hat{\tau}_{OLS} &:= \hat{c}_{(1)} - \hat{c}_{(0)} + \bar{X}(\hat{\beta}_{(1)} - \hat{\beta}_{(0)}) \\ &= \frac{1}{n}\sum_{i=1}^{n}\left((\hat{c}_{(1)} + X_i\hat{\beta}_{(1)}) - (\hat{c}_{(0)} - X_i\hat{\beta}_{(0)})\right),\end{aligned}$$

where $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ and the estimators are obtained by OLS for the two linear models.

Randomized trials in the linear model

OLS estimator

$$\begin{aligned}\hat{\tau}_{OLS} &:= \hat{c}_{(1)} - \hat{c}_{(0)} + \bar{X}(\hat{\beta}_{(1)} - \hat{\beta}_{(0)}) \\ &= \frac{1}{n}\sum_{i=1}^{n}\left((\hat{c}_{(1)} + X_i\hat{\beta}_{(1)}) - (\hat{c}_{(0)} - X_i\hat{\beta}_{(0)})\right),\end{aligned}$$

where $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ and the estimators are obtained by OLS for the two linear models.

Properties of $\hat{\tau}_{OLS}$

- Asymptotic independence of $\hat{c}_{(w)}$, $\hat{\beta}_{(w)}$ and $\bar{X}$ and also

  $$\hat{\tau}_{OLS} - \tau = (\hat{c}_{(1)} - c_{(1)}) - (\hat{c}_{(0)} - c_{(0)}) + \bar{X}(\beta_{(1)} - \beta_{(0)}) + \bar{X}(\hat{\beta}_{(1)} - \hat{\beta}_{(0)} - \beta_{(1)} + \beta_{(0)}).$$

- Noting $V_{OLS} = 4\sigma^2 + (\beta_{(0)} - \beta_{(1)})^T Var(X)(\beta_{(0)} - \beta_{(1)})$, by central limit theorem we get

  $$\sqrt{n}\,(\hat{\tau}_{OLS} - \tau) \xrightarrow[n\to\infty]{d} \mathcal{N}(0, V_{OLS}).$$

Randomized trials in the linear model

Properties of $\hat{\tau}_{OLS}$

- Noting $V_{OLS} = 4\sigma^2 + \|\beta_{(0)} - \beta_{(1)}\|_A^2$, by central limit theorem we get

$$\sqrt{n}\,(\hat{\tau}_{OLS} - \tau) \xrightarrow[n\to\infty]{d} \mathcal{N}(0, V_{OLS}).$$

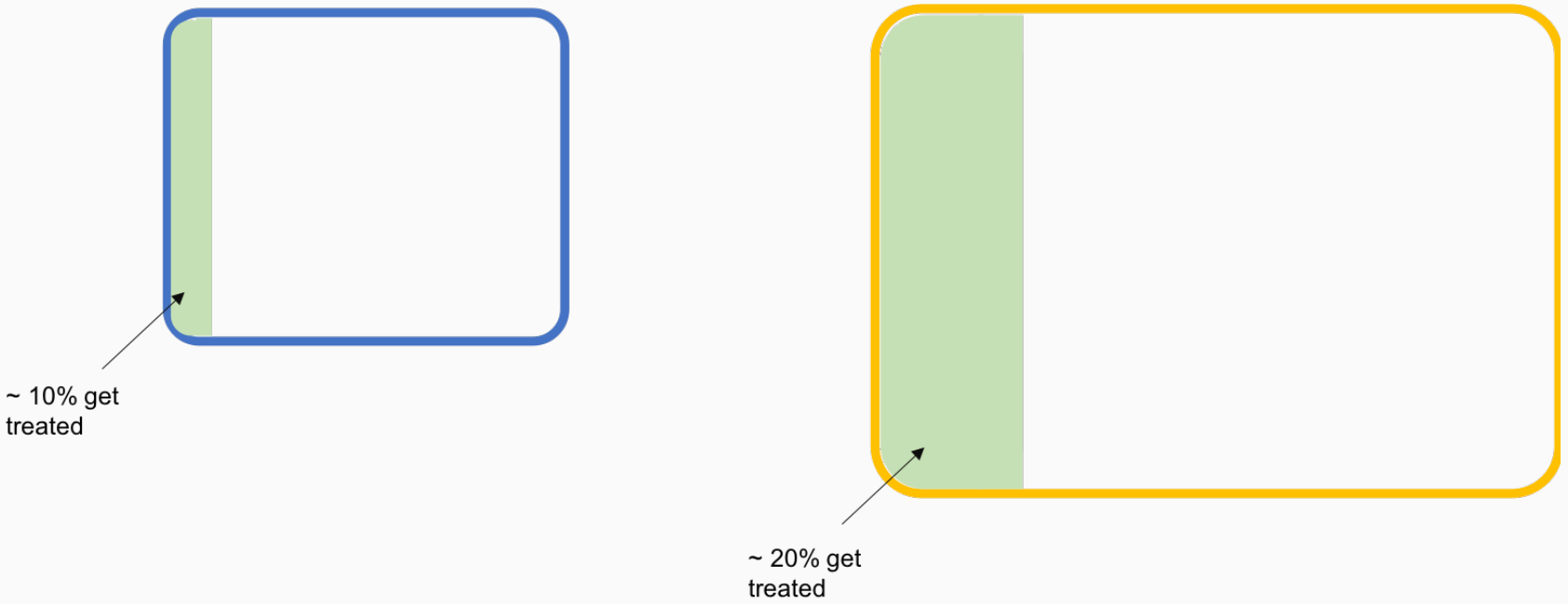
Remark

- Under the linearity assumption,
  $V_{DM} = 4\sigma^2 + \|\beta_{(0)} - \beta_{(1)}\|_A^2 + \|\beta_{(0)} + \beta_{(1)})\|_A^2$.
  $\Rightarrow \hat{\tau}_{OLS}$ is always at least as good as $\hat{\tau}_{DM}$ in terms of asymptotic variance.
- This still holds in case of model mis-specification. *(proof uses Huber-White linear regression analysis)*

Beyond a single randomized controlled trial

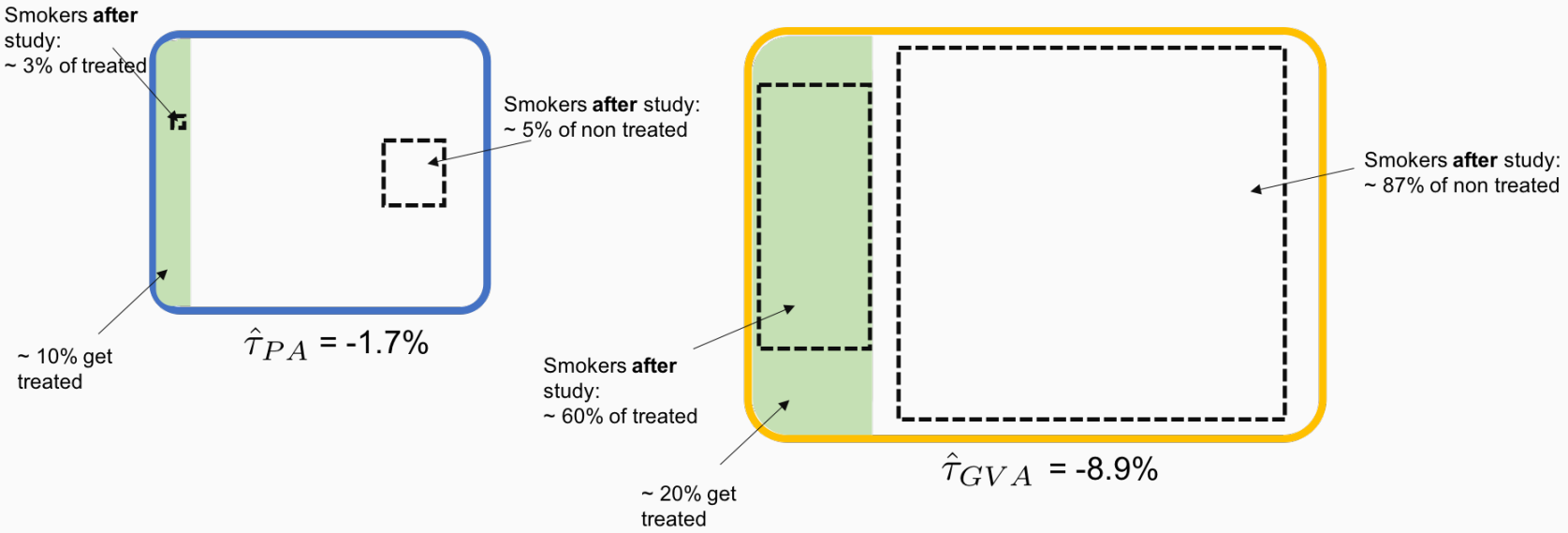
How to combine different experiments or data sets

Study the effect of a cash incentive to discourage teenagers from smoking in two different cities.



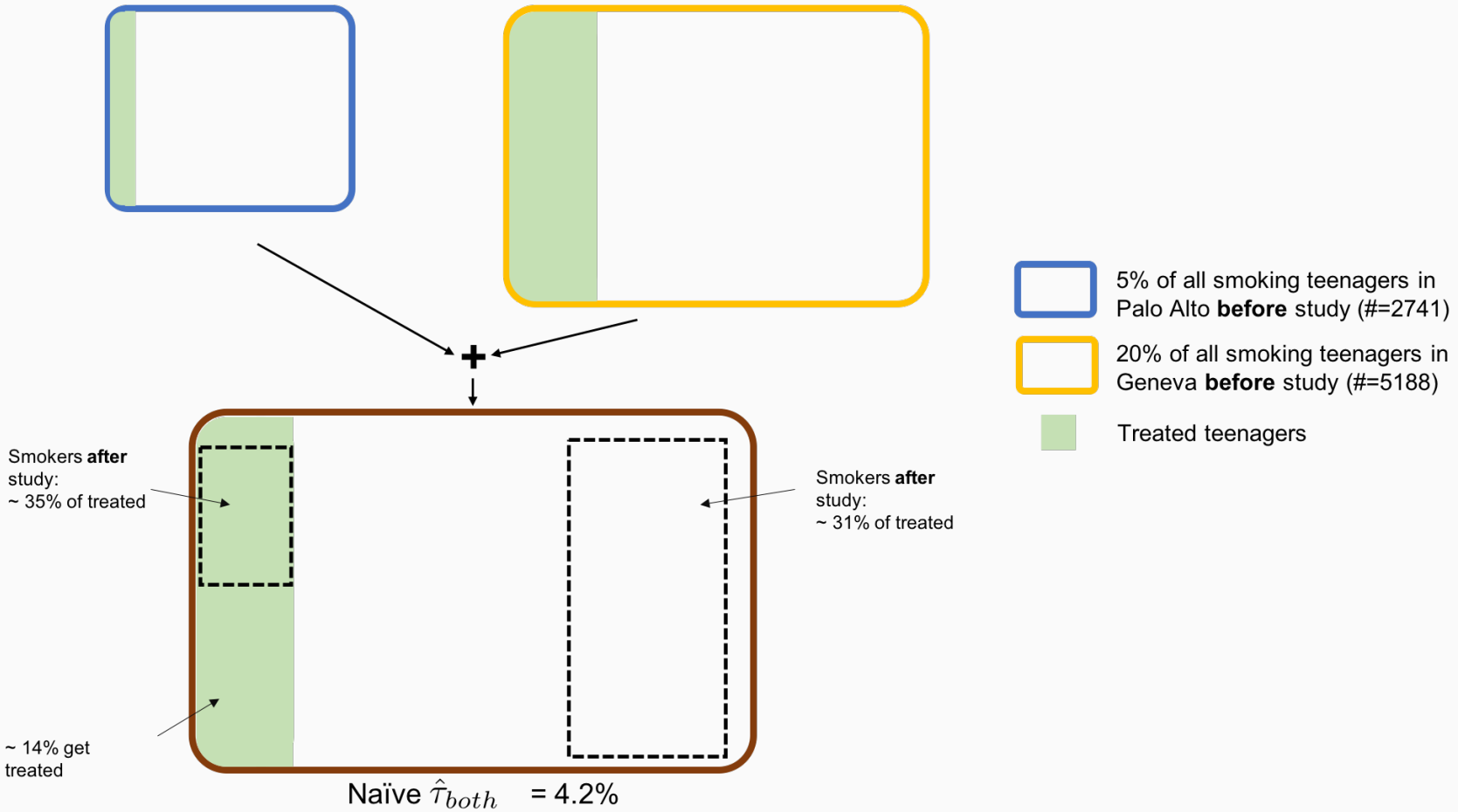
How to combine different experiments or data sets

Study the effect of a cash incentive to discourage teenagers from smoking in two different cities.



How to combine different experiments or data sets

Study the effect of a cash incentive to discourage teenagers from smoking in two different cities.



How to combine different experiments or data sets

Study the effect of a cash incentive to discourage teenagers from smoking in two different cities.

Correct aggregation of the two studies:

$$\hat{\tau}_{both} = \frac{\square}{\square}\,\hat{\tau}_{PA} + \frac{\square}{\square}\,\hat{\tau}_{GVA} = \text{-}6.5\%$$

Aggregating several ATE estimators

How to combine several trials testing the same treatment but on different populations?

Assumptions

- n iid samples (X_i, Y_i, W_i),
- Covariates X_i take values in a **finite discrete** space $\mathcal{X}$ (i.e. $|\mathcal{X}| = p$).
- Treatment assignment is random conditionally on X_i:

$$\{Y_i(0), Y_i(1)\} \perp\!\!\!\perp W_i \,|\, X_i = x, \quad \forall x \in \mathcal{X}.$$

Bucket-wise ATE

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

Results for aggregated difference-in-means estimators

Aggregated difference-in-means estimator

$$\hat{\tau} := \sum_{x \in \mathcal{X}} \frac{n_x}{n} \hat{\tau}(x) = \sum_{x \in \mathcal{X}} \frac{n_x}{n} \left(\frac{1}{n_{x1}} \sum_{\{X_i = x,\, W_i = 1\}} Y_i - \frac{1}{n_{x0}} \sum_{\{X_i = x,\, W_i = 0\}} Y_i \right)$$

- Denoting $e(x) = \mathbb{P}(W_i = 1 \,|\, X_i = x)$ and adding simplifying assumption $Var(Y(w) \,|\, X = x) = \sigma^2(x)$ we can show that

$$\sqrt{n_x} \left(\hat{\tau}(x) - \tau(x) \right) \xrightarrow[n \to \infty]{d} \mathcal{N} \left(0, \frac{\sigma^2(x)}{e(x)(1 - e(x))} \right)$$

- Finally, denoting $V_{BUCKET} = Var(\tau(X)) + \mathbb{E}\left[\frac{\sigma^2(X)}{e(X)(1-e(X))} \right]$,

$$\sqrt{n} \left(\hat{\tau} - \tau \right) \xrightarrow[n \to \infty]{d} \mathcal{N}(0, V_{BUCKET})$$

no dependence in p, # of buckets!

Inverse-propensity weighting

Continuous X and the propensity score

Observation from discrete $\mathcal{X}$ with finite number of buckets: the number of buckets p does not affect the accuracy of inference.

How to transpose the analysis and results to the continuous case?

1. Modify assumptions
2. Define analogue of "buckets"

Assumptions

- n iid samples (X_i, Y_i, W_i),
- Covariates X_i take values in a **continuous** space $\mathcal{X}$.
- Treatment assignment is random conditionally on X_i:

$$\{Y_i(0), Y_i(1)\} \perp\!\!\!\perp W_i \,|\, X_i \quad \equiv \textbf{unconfoundedness} \text{ assumption.}$$

Unconfoundedness and the propensity score

Observation from discrete $\mathcal{X}$ with finite number of buckets: the number of buckets p does not affect the accuracy of inference.

How to transpose the analysis and results to the continuous case?

1. Modify assumpions
2. Define analogue of "buckets"

Propensity score

$$e(x) = \mathbb{P}(W_i = 1 \,|\, X_i = x) \quad \forall\, x \in \mathcal{X}.$$

Unconfoundedness and the propensity score

Propensity score

$$e(x) = \mathbb{P}(W_i = 1 \,|\, X_i = x) \quad \forall\, x \in \mathcal{X}.$$

Key property

e is a balancing score, i.e. under unconfoundedness, it satisfies

$$\{Y_i(0), Y_i(1)\} \perp\!\!\!\perp W_i \,|\, e(X_i)$$

As a consequence, it suffices to control for $e(X)$ (rather than X), to remove biases associated with non-random treatment assignment.

Unconfoundedness and the propensity score: finite number of strata

If the data falls in J strata $(S_j)_{1 \leq j \leq J}$, with $J < \infty$ and such that $e(x) = e_j$ in each stratum, then we have a consistent estimator for ATE:

$$\hat{\tau} := \sum_{j=1}^{J} \frac{n_j}{n} \hat{\tau}_j = \sum_{j=1}^{J} \frac{n_j}{n} \left(\frac{1}{n_{j1}} \sum_{\{X_i \in S_j,\, W_i = 1\}} Y_i - \frac{1}{n_{j0}} \sum_{\{X_i \in S_j,\, W_i = 0\}} Y_i \right)$$

Unconfoundedness and the propensity score: inverse-propensity weighting

The previous finite number of strata assumption is unrealistic. But we can generalize the previous estimator using **propensity score estimates**:

$$\begin{aligned}
\hat{\tau} &:= \sum_{j=1}^{J} \frac{n_j}{n} \left(\frac{1}{n_{j1}} \sum_{\{X_i \in S_j,\, W_i = 1\}} Y_i - \frac{1}{n_{j0}} \sum_{\{X_i \in S_j,\, W_i = 0\}} Y_i \right) \\
&= \frac{1}{n} \sum_{j=1}^{J} \left(\frac{1}{\hat{e}_j} \sum_{\{X_i \in S_j,\, W_i = 1\}} Y_i - \frac{1}{1 - \hat{e}_j} \sum_{\{X_i \in S_j,\, W_i = 0\}} Y_i \right) \\
&= \frac{1}{n} \sum_{i=1}^{n} \left(\frac{W_i Y_i}{\hat{e}(X_i)} - \frac{(1 - W_i) Y_i}{1 - \hat{e}(X_i)} \right).
\end{aligned}$$

here we have $\hat{e}(x) = \hat{e}_j = \frac{n_{j1}}{n_j}$ for all $x \in S_j$ but we could use any other method to estimate $\hat{e}$.

Unconfoundedness and the propensity score: inverse-propensity weighting

And define

$$\hat{\tau}_{IPW} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{W_i Y_i}{\hat{e}(X_i)} - \frac{(1-W_i)Y_i}{1-\hat{e}(X_i)}\right)$$

an inverse-propensity weighted estimation of ATE.

The quality of this estimator depends on the estimation quality of $\hat{e}(x)$.

Propensity score estimation and inverse-propensity weighting

Assume a linear-logistic model:

1. $e(x) = \mathbb{P}(W_i = 1 \,|\, X_i = x) = \frac{1}{1+e^{-x^T\alpha}}$
2. $\mu_{(w)}(x) = x^T\beta_{(w)}$ (for $w \in \{0, 1\}$).
3. $Y_i(w) = \mu_{(W_i)}(X_i) + \varepsilon_i$.

Decompose the general ATE estimator

$$\hat{\tau} = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{\gamma}_{(1)}(X_i)W_iY_i - \hat{\gamma}_{(0)}(X_i)(1-W_i)Y_i\right)$$

as follows:

$$\begin{aligned}\hat{\tau} = &\bar{X}(\beta_{(1)} - \beta_{(0)}) + [\text{term to pay that depends on the noise } \varepsilon] \\ &+ \left(\frac{1}{n}\sum_{i=1}^{n}\hat{\gamma}_{(1)}(X_i)W_iX_i - \bar{X}\right)\beta_{(1)} \\ &- \left(\frac{1}{n}\sum_{i=1}^{n}\hat{\gamma}_{(0)}(X_i)(1-W_i)X_i - \bar{X}\right)\beta_{(0)}\end{aligned}$$

Propensity score estimation and inverse-propensity weighting

Covariate balancing propensity score (CBPS)

- Use $\hat{\gamma}_{(1)} = \frac{1}{\hat{e}(x)} = 1 + e^{-x^T \hat{\alpha}_{(1)}}$ and solve for $\alpha_{(1)}$ by moment matching:

$$\frac{1}{n} \sum_{i=1}^{n} \hat{\gamma}_{(1)}(X_i) W_i X_i - \bar{X} = 0$$

- Same for $\hat{\gamma}_{(0)} = \frac{1}{1-\hat{e}(x)} = \frac{e^{-x^T \hat{\alpha}_{(0)}}}{1+e^{-x^T \hat{\alpha}_{(0)}}}$.

Note that $\hat{\gamma}_{(1)}$ and $\hat{\gamma}_{(0)}$ do not use the same propensity model but we can verify that both $\hat{\alpha}_{(1)}$ and $\hat{\alpha}_{(0)}$ are $\sqrt{n}$-consistent:

$$\|\hat{\alpha}_{(w)} - \alpha\|_2 = \mathcal{O}_P\left(\frac{1}{\sqrt{n}}\right) \qquad \text{for } w \in \{0, 1\}$$

Propensity score estimation and inverse-propensity weighting

IPW with covariate balancing propensity score (CBPS)

Under regularity assumptions (including overlap, i.e. $\exists \eta > 0$ such that $\eta \leq e(x) \leq 1 - \eta$ for all $x \in \mathcal{X}$), we have:

$$\hat{\tau}_{CBPS} = \bar{X}(\beta_{(1)} - \beta_{(0)}) + \frac{1}{n} \sum_{i=1}^{n} \left(\frac{W_i \varepsilon_i}{\hat{e}(X_i)} - \frac{(1 - W_i)\varepsilon_i}{1 - \hat{e}(X_i)} \right) + \mathcal{O}_P \left(\frac{1}{n} \right)$$

And this estimator has same asymptotic variance as for bucketing.

Double robustness property

Double robustness of CBPS

Under linear-logistic specification, $\hat{\tau}_{CBPS}$ has "good" asymptotic variance. What happens if the model is mis-specified?

Double robustness

$\hat{\tau}_{CBPS}$ remains consistent in either one of the following cases:

1. Outcome model is linear but propensity score $e(x)$ is not logistic.
2. Propensity score $e(x)$ is logistic but outcome model is not linear.

Note that the asymptotic variance might be different in these cases.

Another doubly robust ATE estimator

Define $\mu_{(w)}(x) := \mathbb{E}[Y_i(w) \mid X_i = x]$ and $e(x) := \mathbb{P}(W_i = 1 \mid X_i = x)$.

Doubly robust estimator

Assume we have access to estimates $\hat{\mu}_{(w)}$ and $\hat{e}(x)$.

$$\hat{\tau}_{DR} := \frac{1}{n} \sum_{i=1}^{n} \left(\hat{\mu}_{(1)}(X_i) - \hat{\mu}_{(0)}(X_i) + W_i \frac{Y_i - \hat{\mu}_{(1)}(X_i)}{\hat{e}(X_i)} - (1 - W_i) \frac{Y_i - \hat{\mu}_{(0)}(X_i)}{1 - \hat{e}(X_i)} \right)$$

is consistent if either the $\hat{\mu}_{(w)}(x)$ are consistent or $\hat{e}(x)$ is consistent.

Furthermore $\hat{\tau}_{DR^*}$ has same asymptotic variance as $\hat{\tau}_{BUCKET}$ and $\hat{\tau}_{CBPS}$.

Remark: In case of overparametrization or non-parametric estimation $\hat{\mu}_{(w)}(x)$ and $\hat{e}(x)$ should be learned/estimated by cross-validation to avoid overfitting.

Semiparametric efficiency for ATE estimation

Efficient score estimator

Given unconfoundedness ($\{Y_i(1), Y_i(1)\} \perp\!\!\!\perp W_i \,|\, X_i$) but no further parametric assumptions on $\mu_{(w)}(x)$ and $e(x)$, the previously attained asymptotic variance,

$$V^* := Var(\tau(X)) + \mathbb{E}\left[\frac{\sigma^2(X)}{e(X)(1-e(X))}\right],$$

is optimal and any estimator τ^* that attains it is asymptotically equivalent to $\hat{\tau}_{DR^*}$.

V^* is the semiparametric efficient variance for ATE estimation.

Cross-fitting and machine learning for ATE estimation

Cross-fitting for ATE estimation

Cross-fitted ATE estimator

Assume we divide the data into K folds.

$$\hat{\tau}_{CF} = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{\mu}_{(1)}^{(-k(i))}(X_i) - \hat{\mu}_{(0)}^{(-k(i))}(X_i) + W_i\frac{Y_i - \hat{\mu}_{(1)}^{(-k(i))}(X_i)}{\hat{e}^{(-k(i))}(X_i)} - (1-W_i)\frac{Y_i - \hat{\mu}_{(0)}^{(-k(i))}(X_i)}{1-\hat{e}^{(-k(i))}(X_i)}\right),$$

where $k(i)$ maps an observation X_i to one of the K folds and $\hat{\mu}^{(-j)}$ indicates that the estimator has been learned on all the folds except the j-th fold.

Assuming overlap, sup-norm consistency of all used machine learning adjustments and risk decay, we have

$$\sqrt{n}\,(\hat{\tau}_{CF} - \hat{\tau}_{DR^*}) \xrightarrow[n\to\infty]{p} 0.$$

And we can prove that we can build level-α confidence intervals for τ.

Heterogeneous treatment Effect Estimation

Instead of estimating the average treatment effect, we may seek to estimate the conditional average treatment effect function

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

$\rightarrow$ harder to solve this problem (note that $\tau = \mathbb{E}[\tau(X)]$).

$\rightarrow$ take care of regularization bias (different amounts of regularization in treatement and control models).

$\rightarrow$ need for further investigations in different directions.

Optimal policy estimation

Beyond causal inference: based on the heterogeneous treatment effect, establish decision rules by defining an optimal policy:

Given $\Pi = \{\pi : \mathcal{X} \to \{0, 1\}$, with potentially some constraints on $\pi\}$, find the policy that maximizes the expected utility $\mathbb{E}[Y_i(\pi(X_i))]$ or that minimizes the regret.

Conclusion

Summary

Problem and question of interest

- Estimate the effect of a treatment on an individual via a potential outcomes model.
- Inevitably faced with missing values (we only observe one outcome per individual).

Summary

Problem and question of interest

- Estimate the effect of a treatment on an individual via a potential outcomes model.
- Inevitably faced with missing values (we only observe one outcome per individual).

Established approach(es)

- First solution: randomized controlled trials (RCT).
- Second solution: bucketing/inverse-propensity weighting to adjust for biases in the treatment assignment.
  - Efficient score estimator is computationally feasible (by using cross-fitting).
  - Double robustness property for model mis-specifications.
  - Using machine learning approaches do not harm the interpretability of the causal effect estimation.

Objectives for Traumabase and traumatic brain injury (TBI)

- Traumatic brain injury is a very heterogeneous injury: the patients injury/physiological profiles can differ a lot and the symptoms and degrees of severity cover a large spectrum. Is it still possible to estimate causal effects using the potential outcomes model?
  - Does the administration of tranexamic acid have an effect on mortality?
    $\rightarrow$ single treatment and binary outcome, currently studied by student group.
  - Do certain treatment strategies, i.e. bundles of treatments (administration of noradrenaline and SSH and tranexamic acid, etc.), have an effect on 24h mortality, on 14d mortality, etc.?
    $\rightarrow$ more methodological investigations needed to perform causal inference for this type of question.

Alternatives to potential outcomes models.

Potential outcomes model proposed by Neyman (1923) and Rubin (1974). But there are other approaches to causal effect estimation:

- Structural equation models (common in economics, social sciences).
- Instrumental variables (Wright, 1928)

And the causal inference model can be made more rich by introducing "mediators" which are affected by the treatment and linked to the outcome.

Do you have any questions or comments?

References I

- Loftus, J. (2015). Lecture on Causal Inference. Stanford University.
- Pearl, J. (2000). *Causality: Models, Reasoning and Inference*. Cambridge University Press, New York, NY.
- Wager, S. (2018). Lecture Notes on Causal Inference (OIT 661). Stanford University.