CSDS 452 Causality and Machine Learning

Lecture 5: Traditional Causal Effect Estimation Methods

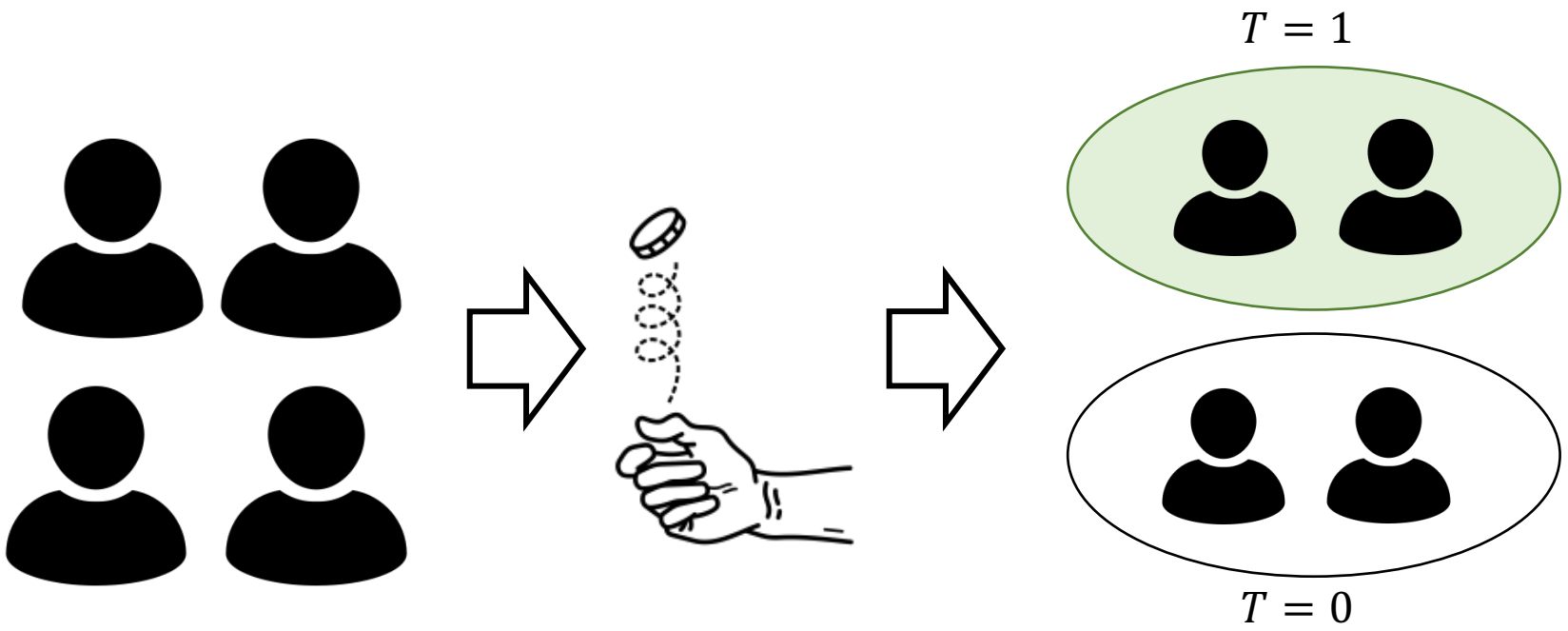
Instructor: Jing Ma

Fall 2024, CDS@CWRU

Outline

- Regression Adjustment
  - Conditional outcome modeling
  - Grouped conditional outcome modeling
  - X-learner
- Re-weighting methods (continue)
  - CBPS
  - Trimming
  - D2VD
- Subspace Learning
- Difference-in-Difference

Recap: Randomized Experiments



- Gold standard to assess the causal effect.
- The allocation of the treatment is under control. The distribution of the covariates for treated and control patients is balanced.

Recap: Backdoor criterion and backdoor adjustment

- A set of variables W satisfies the **backdoor criterion** relative to T and Y if the following are true:
  - W blocks all backdoor paths from T to Y
  - W does not contain any descendants of T

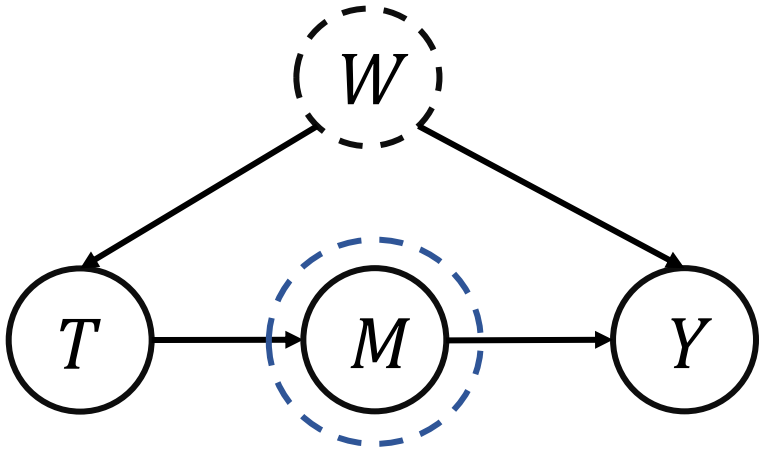
- Given the modularity assumption and that W satisfies the backdoor criterion, we can identify the causal effect of T on Y:

$$P(y|do(t)) = \sum_w P(w)P(y|t,w)$$

Recap: Frontdoor Criterion

- A set of variables M satisfies the frontdoor criterion relative to T and Y if the following are true:
  - 1. M completely mediates the effect of T on Y (i.e., all causal paths from T to Y go through M).
  - There is no unblocked backdoor path from T to M.
  - All backdoor paths from M to Y are blocked by T.

$$P(y|do(t))$$
$$= \sum_m P(m|do(t))P(y|do(m))$$
$$= \sum_m P(m|t) \sum_{t'} P(y|m,t')P(t')$$



Recap: 3 Rules in do-calculus

- **Rule 1:** If $Y \perp\!\!\!\perp_{G_{\overline{T}}} Z \mid T, W$:

$$P(y|do(t), z, w) = P(y|do(t), w)$$

- **Rule 2**: If $Y \perp\!\!\!\perp_{G_{\overline{T},\underline{Z}}} Z \mid T, W$:

$$P(y|do(t), do(z), w) = P(y|do(t), z, w)$$

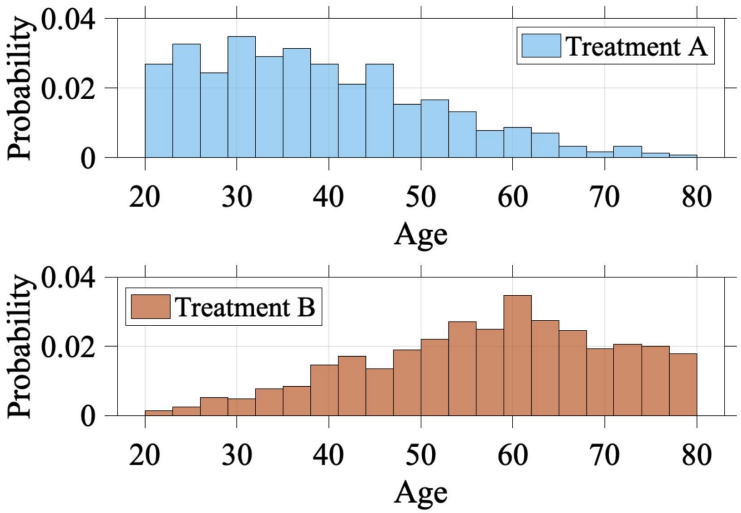
- **Rule 3**: If $Y \perp\!\!\!\perp_{G_{\overline{T},\overline{Z(W)}}} Z \mid T, W$

$$P(y|do(t), do(z), w) = P(y|do(t), w)$$

Identification in SCM: Use the 3 rules, until there is no do-operator.

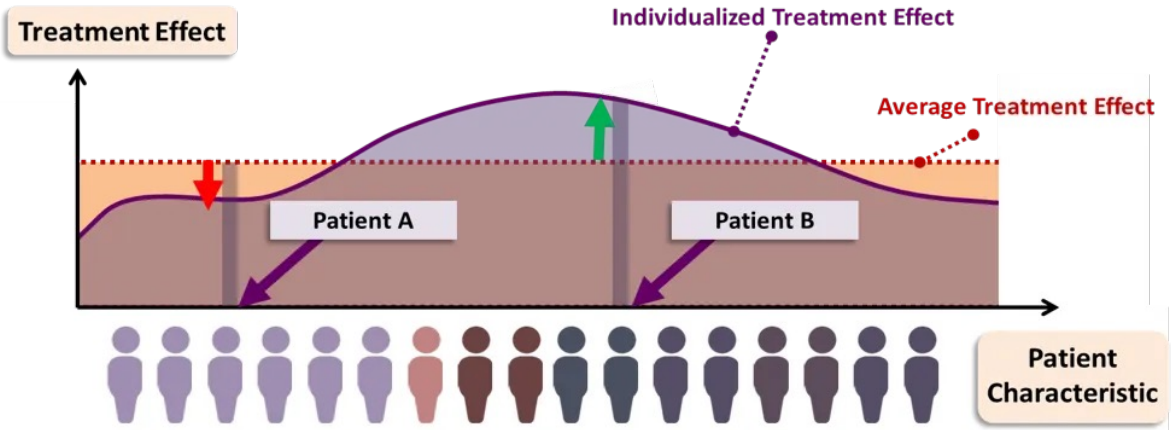
Selection Bias

- **Selection Bias**: The distribution of the observed group is not representative to the group we are interested in
- Confounder variables affect units' treatment choices, which leads to the selection bias
- Selection bias makes counterfactual outcome estimation more difficult



Recap: Treatment Effect

- Individual treatment effect (ITE):
  - ITE $= Y_i(1) - Y_i(0)$
- Average treatment effect (ATE):
  - ATE $= E[Y_i(1) - Y_i(0)]$
- Conditional average treatment effect (CATE):
  - CATE = $E[Y_i(1) - Y_i(0)|X = x]$



https://www.vanderschaar-lab.com/individualized-treatment-effect-inference/

Outline

- Regression Adjustment
  - Conditional outcome modeling
  - Grouped conditional outcome modeling
  - X-learner
- Re-weighting methods (continue)
  - CBPS
  - Trimming
  - D2VD
- Subspace Learning
- Difference-in-Difference

Regression Adjustment

- Conditional average treatment effect (CATE):
  - CATE = $E[Y_i(1) - Y_i(0)|X = x]$
- Based on the potential outcome framework, we infer the counterfactual outcomes $y_i^{1-t_i}$ with the features x and the treatment t

Conditional outcome modeling (COM)

$$\tau = E_W[E[Y|T = 1, W] - E[Y|T = 0, W]]$$

Conditional outcome modeling (COM)

$$\tau = E_W[\underbrace{E[Y|T=1,W]}_{\text{model}} - \underbrace{E[Y|T=0,W]}_{\text{model}}]$$

Conditional outcome modeling (COM)

$$\tau = E_W[E[Y|T=1,W] - E[Y|T=0,W]]$$

model model

$$\tau = E_W[\mu(1,W) - \mu(0,W)]$$

model model

Conditional outcome modeling (COM)

$$\tau = E_W[E[Y|T=1,W] - E[Y|T=0,W]]$$

model model

$$\tau = E_W[\mu(1,W) - \mu(0,W)]$$

model model

Estimator: $\hat{\tau} = \frac{1}{n}\sum_i(\hat{\mu}(1,w_i) - \hat{\mu}(0,w_i))$

COM estimation of CATEs

- ATE COM Estimator: $\hat{\tau} = \frac{1}{n}\sum_i(\hat{\mu}(1, w_i) - \hat{\mu}(0, w_i))$
- CATE Estimand:

$$\tau(x) \triangleq E[Y(1) - Y(0) \,|X = x]$$
$$= E_W[E[Y|T = 1, X = x, W] - E[Y|T = 0, X = x, W]]$$

COM estimation of CATEs

- ATE COM Estimator: $\hat{\tau} = \frac{1}{n}\sum_i(\hat{\mu}(1, w_i) - \hat{\mu}(0, w_i))$
- CATE Estimand:

$$\tau(x) \triangleq E[Y(1) - Y(0) \,|X = x]$$
$$= E_W[E[Y|T = 1, X = x, W] - E[Y|T = 0, X = x, W]]$$

$$\mu(t, w, x) \triangleq E[Y|T = t, W = w, X = x]$$

COM estimation of CATEs

- ATE COM Estimator: $\hat{\tau} = \frac{1}{n}\sum_i(\hat{\mu}(1, w_i) - \hat{\mu}(0, w_i))$
- CATE Estimand:

$$\tau(x) \triangleq E[Y(1) - Y(0) | X = x]$$
$$= E_W[E[Y|T = 1, X = x, W] - E[Y|T = 0, X = x, W]]$$

$$\mu(t, w, x) \triangleq E[Y|T = t, W = w, X = x]$$

- CATE COM Estimator:

$$\hat{\tau}(x) \triangleq \frac{1}{n_x} \sum_{i:x_i=x} (\hat{\mu}(1, w_i, x) - \hat{\mu}(0, w_i, x))$$

COM estimation of CATEs

- ATE COM Estimator: $\hat{\tau} = \frac{1}{n}\sum_i(\hat{\mu}(1, w_i) - \hat{\mu}(0, w_i))$
- CATE Estimand:

$$\tau(x) \triangleq E[Y(1) - Y(0) \,|X = x]$$
$$= E_W[E[Y|T = 1, X = x, W] - E[Y|T = 0, X = x, W]]$$

$$\mu(t, w, x) \triangleq E[Y|T = t, W = w, X = x]$$

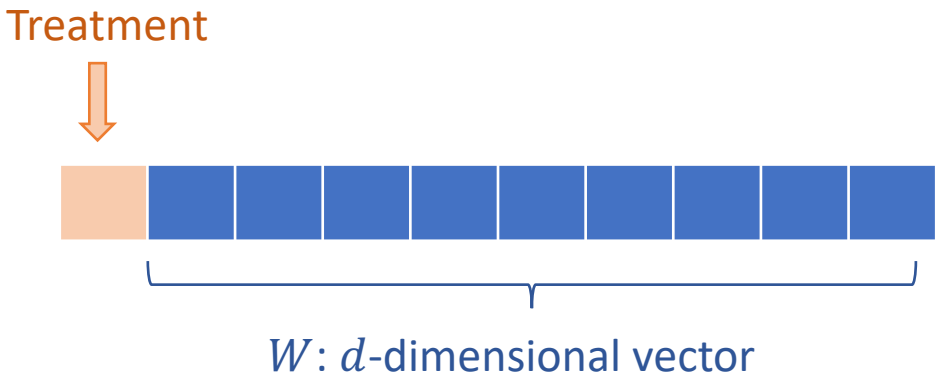
- CATE COM Estimator:

$$\hat{\tau}(x) \triangleq \frac{1}{n_x} \sum_{i:x_i=x} (\hat{\mu}(1, w_i, x) - \hat{\mu}(0, w_i, x))$$

Other names of COM estimator: G-computation estimators, where G is for “generalized”; Parametric G-formula; S-learner where “S” is for “Single”

Problem with COM estimation in high dimensions

- $\hat{\tau} = \frac{1}{n}\sum_i(\hat{\mu}(1, w_i) - \hat{\mu}(0, w_i))$
- In high dimension dataset (suppose W is a d-dimensional vector), the estimator may ignore T and the estimate can be biased toward zero [1]



[1] Künzel S R, et al. Metalearners for estimating heterogeneous treatment effects using machine learning[J]. Proceedings of the national academy of sciences, 2019, 116(10): 4156-4165.

Outline

- Regression Adjustment
  - Conditional outcome modeling
  - Grouped conditional outcome modeling
  - X-learner
- Re-weighting methods (continue)
  - CBPS
  - Trimming
  - D2VD
- Subspace Learning
- Difference-in-Difference

Grouped COM (GCOM) estimation

- Motivation: don't let the model ignore T
- COM Estimator:

$$\hat{\tau} = \frac{1}{n}\sum_{i}(\hat{\mu}(1, w_i) - \hat{\mu}(0, w_i))$$

- GCOM Estimator

$$\hat{\tau} = \frac{1}{n}\sum_{i}(\hat{\mu}_1(w_i) - \hat{\mu}_0(w_i))$$

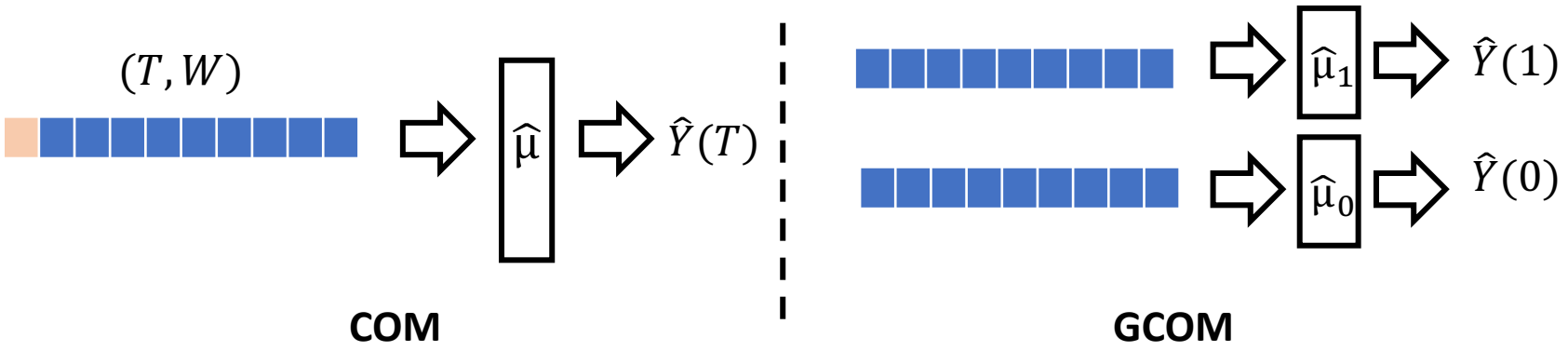
Grouped COM (GCOM) estimation

- Motivation: don't let the model ignore T
- COM Estimator:

$$\hat{\tau} = \frac{1}{n}\sum_i (\hat{\mu}(1, w_i) - \hat{\mu}(0, w_i))$$

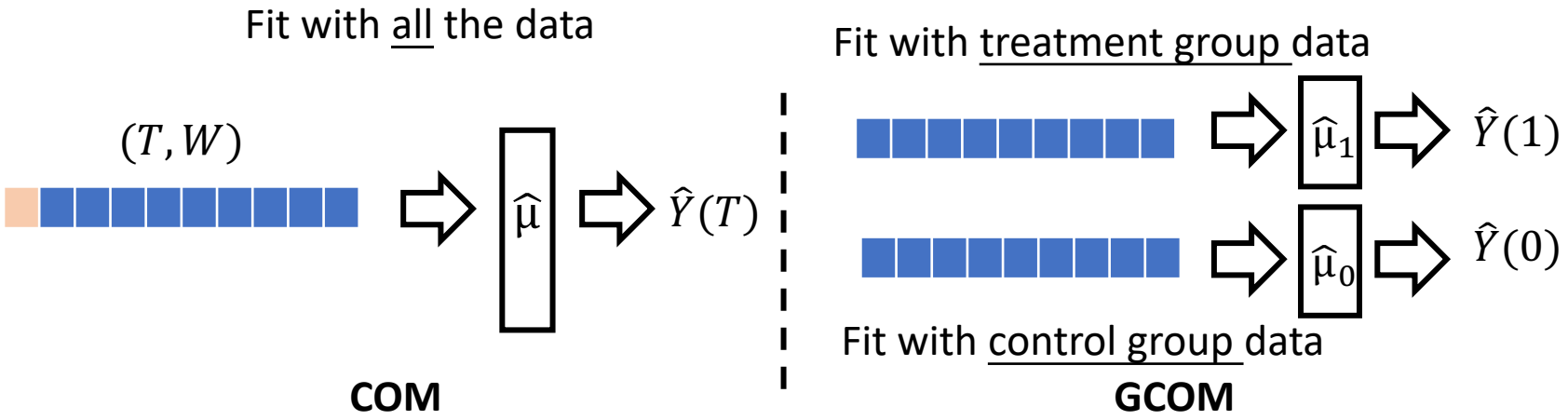
- GCOM Estimator

$$\hat{\tau} = \frac{1}{n}\sum_i (\hat{\mu}_1(w_i) - \hat{\mu}_0(w_i))$$



Grouped COM (GCOM) estimation

- Also known as "T-Learner", "T" for "Two" base models
- Drawback: networks have higher variance than they would if they were trained with all the data



Outline

- Regression Adjustment
  - Conditional outcome modeling
  - Grouped conditional outcome modeling
  - X-learner
- Re-weighting methods (continue)
  - CBPS
  - Trimming
  - D2VD
- Subspace Learning
- Difference-in-Difference

X-Learner

- Limitations of S-Learner and T-Learner:
  - highly rely on the performance of the trained base models, therefore, performance will degrade in very unbalanced data (i.e., the number of one group is much larger than the other).
- X-learner[1] is a method which addresses this issue by adopting information from the control group to give a better estimator on the treated group and vice versa

[1] Künzel S R, et al. Metalearners for estimating heterogeneous treatment effects using machine learning[J]. Proceedings of the national academy of sciences, 2019, 116(10): 4156-4165.

X-Learner

- 1. Estimate $\hat{\mu}_1(x)$ and $\hat{\mu}_0(x)$

Assume X is a sufficient adjustment set and is all observed covariates

Künzel S R, et al. Metalearners for estimating heterogeneous treatment effects using machine learning[J]. Proceedings of the national academy of sciences, 2019, 116(10): 4156-4165.

X-Learner

- 1. Estimate $\hat{\mu}_1(x)$ and $\hat{\mu}_0(x)$
- 2a) Impute ITEs

Assume X is a sufficient adjustment set and is all observed covariates

$$\hat{\tau}_{1,i} = Y_i(1) - \hat{\mu}_0(x_i)$$

Treatment group

$$\hat{\tau}_{0,i} = \hat{\mu}_1(x_i) - Y_i(0)$$

Control group

Künzel S R, et al. Metalearners for estimating heterogeneous treatment effects using machine learning[J]. Proceedings of the national academy of sciences, 2019, 116(10): 4156-4165.

X-Learner

- 1. Estimate $\hat{\mu}_1(x)$ and $\hat{\mu}_0(x)$
- 2a) Impute ITEs

Assume X is a sufficient adjustment set and is all observed covariates

$$\hat{\tau}_{1,i} = Y_i(1) - \hat{\mu}_0(x_i)$$

Treatment group

$$\hat{\tau}_{0,i} = \hat{\mu}_1(x_i) - Y_i(0)$$

Control group

- 2b)
  - Fit a model $\hat{\tau}_1(x)$ to predict $\hat{\tau}_{1,i}$ from x_i in treatment group;
  - Fit a model $\hat{\tau}_0(x)$ to predict $\hat{\tau}_{0,i}$ from x_i in control group

Künzel S R, et al. Metalearners for estimating heterogeneous treatment effects using machine learning[J]. Proceedings of the national academy of sciences, 2019, 116(10): 4156-4165.

X-Learner

- 1. Estimate $\hat{\mu}_1(x)$ and $\hat{\mu}_0(x)$
- 2a) Impute ITEs

Assume X is a sufficient adjustment set and is all observed covariates

$$\hat{\tau}_{1,i} = Y_i(1) - \hat{\mu}_0(x_i)$$

Treatment group

$$\hat{\tau}_{0,i} = \hat{\mu}_1(x_i) - Y_i(0)$$

Control group

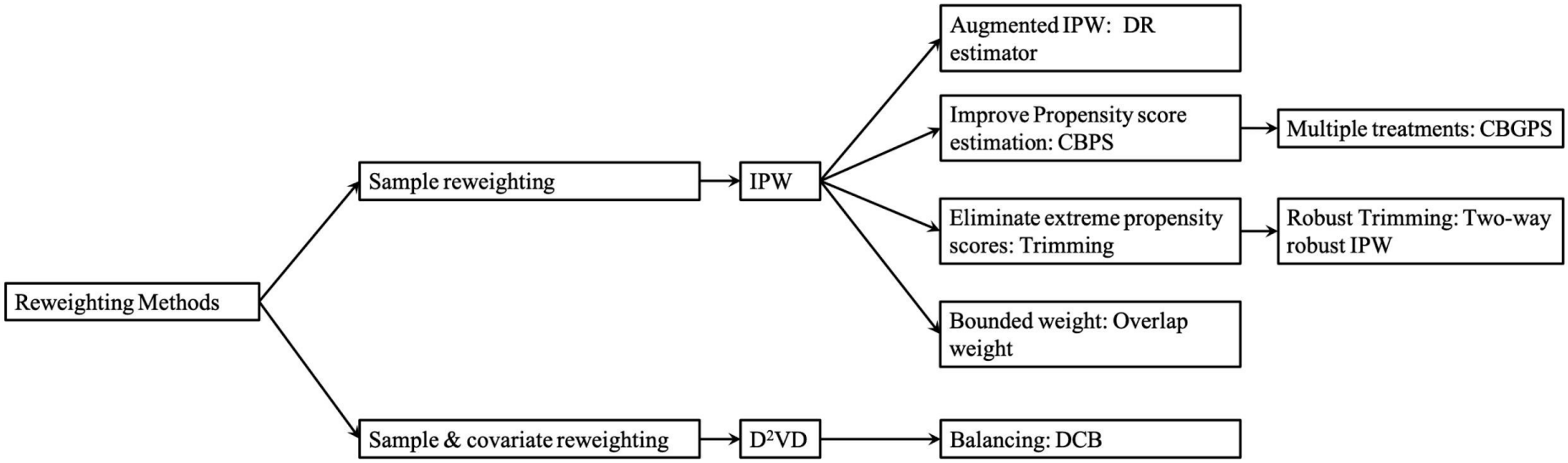
- 2b)
  - Fit a model $\hat{\tau}_1(x)$ to predict $\hat{\tau}_{1,i}$ from x_i in treatment group;
  - Fit a model $\hat{\tau}_0(x)$ to predict $\hat{\tau}_{0,i}$ from x_i in control group
- 3. $\hat{\tau}(x) = g(x)\hat{\tau}_0(x) + (1 - g(x))\hat{\tau}_1(x)$
  - where $g(x)$ is some weighing function between 0 and 1. Example: propensity score

Künzel S R, et al. Metalearners for estimating heterogeneous treatment effects using machine learning[J]. Proceedings of the national academy of sciences, 2019, 116(10): 4156-4165.

Outline

- Regression Adjustment
  - Conditional outcome modeling
  - Grouped conditional outcome modeling
  - X-learner
- Re-weighting methods (continue)
  - CBPS
  - Trimming
  - D2VD
- Subspace Learning
- Difference-in-Difference

Summary of Re-weighting Methods



Yao L, Chu Z, Li S, et al. A survey on causal inference[J]. ACM Transactions on Knowledge Discovery from Data (TKDD), 2021, 15(5): 1-46.

Propensity Score Theorem

- Given positivity, unconfoundedness given W implies unconfoundedness given the propensity score $e(W)$.

$$\big(Y(1), Y(0)\big) \perp\!\!\!\perp T \,|W \;\Rightarrow (Y(1), Y(0)) \perp\!\!\!\perp T \,|\, e(W)$$

Implications for the Positivity-Unconfoundedness Tradeoff

- Recall: overlap decreases with the dimensionality of the adjustment set
- The propensity score magically reduces the dimensionality of the adjustment set done to 1 (a scalar), therefore, it is less likely to violate positivity.

- Unfortunately, we don't directly have it. The best we can do is model it.

Recap: Sample Re-weighting Methods

- **Inverse propensity weighting (IPW)**
- The weight assigned for each unit is:

$$r = \frac{W}{e(x)} + \frac{1-W}{1-e(x)}$$

$W = 1 \Rightarrow r = \frac{1}{e(x)}$

$W = 0 \Rightarrow r = \frac{1}{1-e(x)}$

  where W is the treatment and $e(x)$ is the propensity score

- After re-weighting, the IPW estimator of ATE is defined as:

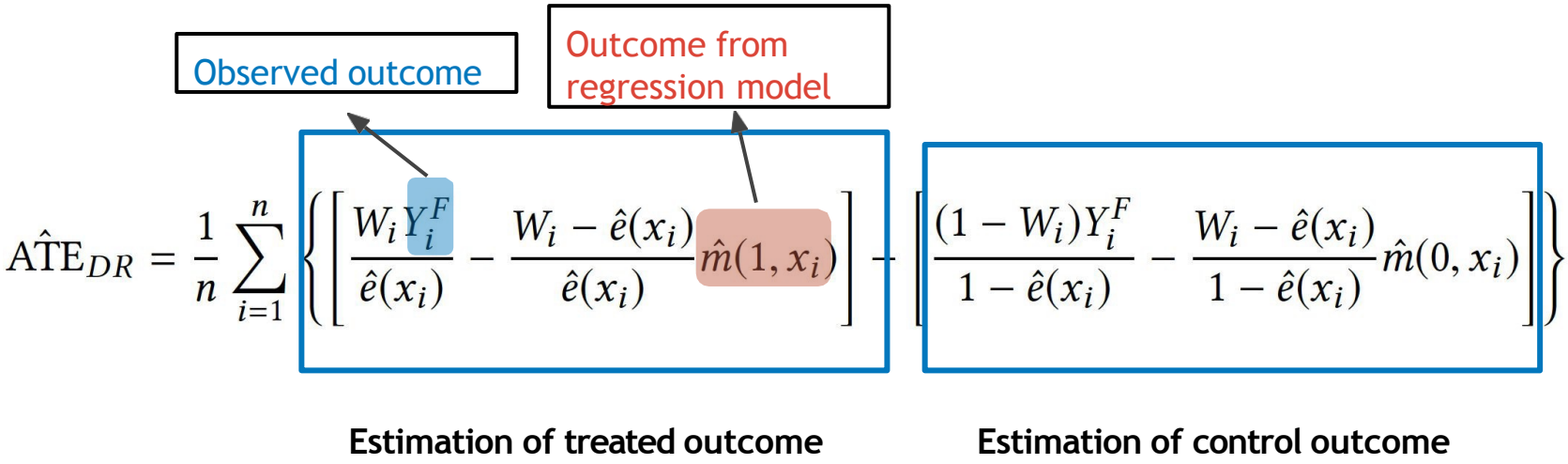
$$\hat{\mathrm{ATE}}_{IPW} = \frac{1}{n}\sum_{i=1}^{n}\frac{W_i Y_i^F}{\hat{e}(x)} - \frac{1}{n}\sum_{i=1}^{n}\frac{(1-W_i)Y_i^F}{1-\hat{e}(x)}$$

  - Theoretical results show that adjustment for the scalar propensity score is sufficient to remove bias due to all observed covariates.
  - However, IPW highly relies on the correctness of propensity scores

J. Robins, A. Rotnitzky, and L. Zhao. "Estimation of regression coefficients when some regressors are not always observed." *Journal of the American statistical Association* 89.427 (1994): 846-866.

Recap: Sample Re-weighting Methods

- **Doubly Robust Estimator (DR)** or **Augmented IPW**
  - Unbiased when one of the propensity score or outcome regression is correct



J. Robins, A. Rotnitzky, and L. Zhao. "Estimation of regression coefficients when some regressors are not always observed." *Journal of the American statistical Association* 89.427 (1994): 846-866.

Outline

- Regression Adjustment
  - Conditional outcome modeling
  - Grouped conditional outcome modeling
  - X-learner
- Re-weighting methods (continue)
  - CBPS
  - Trimming
  - D2VD
- Subspace Learning
- Difference-in-Difference

Covariate balancing propensity score (CBPS)

- Limitation of IPW: highly rely on the propensity score estimation
  - Doubly robust estimator: consults outcomes to make the IPW estimator robust, even when propensity score is not correct.
- An alternative way is to improve the estimation of propensity scores
  - However, propensity scores are easily incorrectly estimated, especially for high-dimensional covariates

Covariate balancing propensity score (CBPS)

- Propensity score tautology: The estimated propensity score is appropriate if it balances covariates
- CBPS exploit the dual characteristics of the propensity score as a covariate balancing score and the conditional probability of treatment assignment.

K. Imai and M. Ratkovic. Covariate balancing propensity score. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 76(1):243–263. 2014

Covariate balancing propensity score (CBPS)

- CBPS models propensity scores and balance covariate simultaneously
- CBPS estimates the propensity score by solving the following problem:

$\tilde{x}_i = f(x_i)$ is a predefined vector-valued measurable function specified by researcher

$$\mathbb{E}\left[\frac{W_i \tilde{x}_i}{e(x_i;\beta)} - \frac{(1-W_i)\tilde{x}_i}{1-e(x_i;\beta)}\right] = 0$$

β: learnable parameters, e.g., logistic model parameters

- For example, by setting $\tilde{x}_i = x_i$, we can ensure that the first moment of each covariate is balanced even when the model is misspecified.

K. Imai and M. Ratkovic. Covariate balancing propensity score. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 76(1):243–263. 2014

Covariate balancing propensity score (CBPS)

- CBPS constructs the covariate balancing score from the estimated parametric propensity score, increasing the robustness to mild misspecification of the propensity score model

Outline

- Regression Adjustment
  - Conditional outcome modeling
  - Grouped conditional outcome modeling
  - X-learner
- Re-weighting methods (continue)
  - CBPS
  - Trimming
  - D2VD
- Subspace Learning
- Difference-in-Difference

Trimming

- Drawback of IPW: unstable if the estimated propensity scores are small

$$r = \frac{W}{e(x)} + \frac{1-W}{1-e(x)}$$

When denominator is close to 0, weight be extremely large

- **Trimming**: routinely employed as a regularization strategy, which eliminates the samples whose propensity scores are less than a pre-defined threshold
  - Drawback: highly sensitive to the amount of trimming
  - **Robust trimming [1]**: combines subsampling with a local polynomial regression based trimming bias corrector, so that it is robust to both small propensity score and the large scale of trimming threshold.

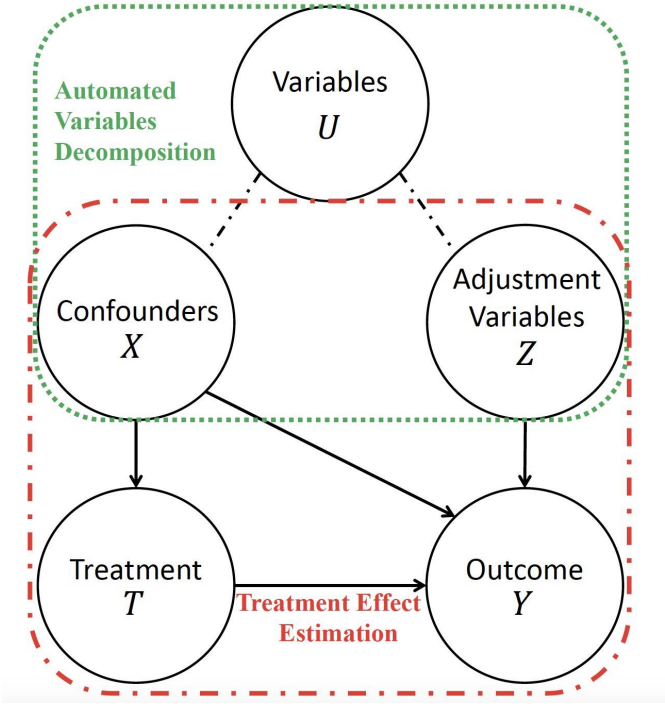
[1] X. Ma and J. Wang. Robust inference using inverse probability weighting. Journal of the American Statistical Association, pages 1–10, 2010.

Outline

- Regression Adjustment
  - Conditional outcome modeling
  - Grouped conditional outcome modeling
  - X-learner
- Re-weighting methods (continue)
  - CBPS
  - Trimming
  - D2VD
- Subspace Learning
- Difference-in-Difference

Data-Driven Variable Decomposition (D2VD)

- Assumption: Not all observed variables are confounders. Observed variables can be decomposed into
  - confounders
  - adjusted variables
  - irrelevant variables
- D^{2}VD distinguishes the confounders and adjustment variables, and meanwhile, eliminates the irrelevant variables.



Kuang, Kun, et al. "Treatment effect estimation with data-driven variable decomposition." *AAAI'17*.

Data-Driven Variable Decomposition (D2VD)

- In IPW, we can consider a transformed outcome as follows:

$$Y^{\star} = Y^{obs} \cdot \frac{T - e(\mathbf{U})}{e(\mathbf{U}) \cdot (1 - e(\mathbf{U}))} = Y^{obs} \cdot \frac{T - e(\mathbf{X})}{e(\mathbf{X}) \cdot (1 - e(\mathbf{X}))}$$

When T=1, $Y^{*} = \frac{Y^{obs}}{e(X)}$

When T=0, $Y^{*} = -\frac{Y^{obs}}{1 - e(X)}$

- Then we have

$$\widehat{ATE}_{IPW} = \widehat{E}(Y^{\star}) = \widehat{E}\left(Y^{obs} \cdot \frac{T - e(\mathbf{X})}{e(\mathbf{X}) \cdot (1 - e(\mathbf{X}))}\right)$$

[1] X. Ma and J. Wang. Robust inference using inverse probability weighting. Journal of the American Statistical Association, pages 1–10, 2010.

Data-Driven Variable Decomposition (D2VD)

- In D2VD, it defines an adjusted transformed outcome

$$Y^{+} = \left(Y^{obs} - \phi(\mathbf{Z})\right) \cdot \frac{T - e(\mathbf{X})}{e(\mathbf{X}) \cdot (1 - e(\mathbf{X}))}$$

where ϕ(Z) helps to reduce the variance among Y , which are associated with Z

- A new adjusted ATE estimator is proposed

$$\widehat{ATE}_{adj} = \widehat{E}(Y^{+}) = \widehat{E}\left(\left(Y^{obs} - \phi(\mathbf{Z})\right) \cdot \frac{T - e(\mathbf{X})}{e(\mathbf{X}) \cdot (1 - e(\mathbf{X}))}\right)$$

Based on proof in [1], the adjusted estimator of ATE is unbiased, and the variance is no greater than IPW

[1] Kuang, Kun, et al. "Treatment effect estimation with data-driven variable decomposition." *AAAI'17*.

Data-Driven Variable Decomposition (D2VD)

How to automatically separate the adjustment variables and confounders?

$$
\begin{aligned}
minimize \quad & \|(Y^{obs} - \mathbf{U}\alpha) \odot W(\beta) - \mathbf{U}\gamma\|_2^2, \\
s.t. \quad & \sum_{i=1}^{m} \log(1 + \exp((1 - 2T_i) \cdot U_i\beta)) < \tau, \\
& \|\alpha\|_1 \leq \lambda,\ \|\beta\|_1 \leq \delta,\ \|\gamma\|_1 \leq \eta,\ \|\alpha \odot \beta\|_2^2 = 0
\end{aligned}
$$

Kuang, Kun, et al. "Treatment effect estimation with data-driven variable decomposition." *AAAI'17*.

Data-Driven Variable Decomposition (D2VD)

Objective: a l2-loss between adjusted outcome and linear regression function on all observed variables

separating confounders X

separating the adjustment variables Z

$$minimize \quad \|(Y^{obs} - \mathbf{U}\alpha) \odot W(\beta) - \mathbf{U}\gamma\|_2^2,$$

ATE estimation function

Adjusted ATE

$$s.t. \quad \sum_{i=1}^{m} \log(1 + \exp((1 - 2T_i) \cdot U_i\beta)) < \tau,$$

Propensity score estimation loss

$$\|\alpha\|_1 \leq \lambda, \ \|\beta\|_1 \leq \delta, \ \|\gamma\|_1 \leq \eta, \ \|\alpha \odot \beta\|_2^2 = 0$$

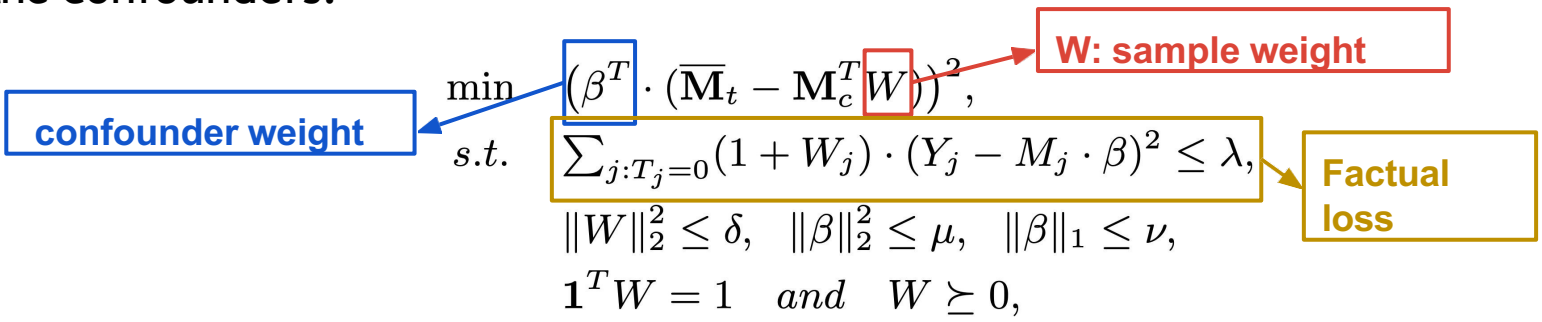
Hadamard product ensures the separation of Z and X

Regularization to distinguish confounder and adjusted variables

Kuang, Kun, et al. "Treatment effect estimation with data-driven variable decomposition." *AAAI'17*.

Sample & Covariate Re-weighting

- **Differentiated Confounder Balancing (DCB)**
- DCB selects, differentiates confounders to balance the distribution.
- DCB balances the distribution by re-weighting both the sample and the confounders.



Kuang, Kun, et al. "Estimating treatment effect in the wild via differentiated confounder balancing." *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2017.

Outline

- Regression Adjustment
  - Conditional outcome modeling
  - Grouped conditional outcome modeling
  - X-learner
- Re-weighting methods (continue)
  - CBPS
  - Trimming
  - D2VD
- Subspace Learning
- Difference-in-Difference

Subspace Learning

- **Goal:** Learning low-dimensional subspaces for dimensionality reduction
- Representative subspace learning methods include: principal component analysis (PCA), locality preserving projections (LPP), canonical correlation analysis (CCA), etc.

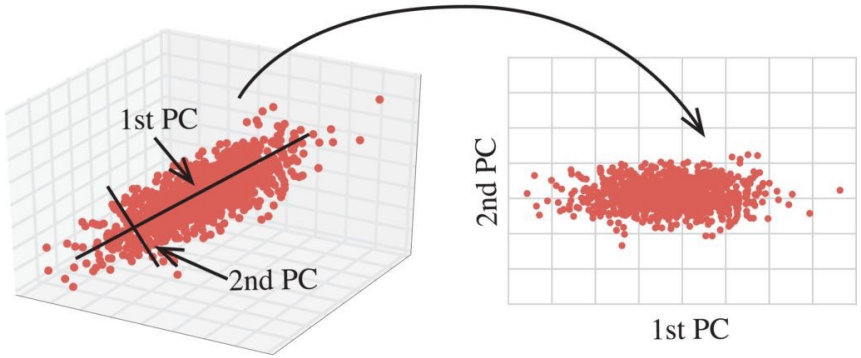


Figure: https://medium.com/@TheDataGyan/dimensionality-reduction-with-pca-and-t-sne-in-r-2715683819

Subspace Learning for Causal Inference

- **Motivation**: Matching in the original data space is simple and flexible, but it could be misled by variables that do not affect the outcome. To address this issue, matching could be performed in subspaces instead.
- **Methods**
  - Random Subspaces
  - Informative Subspace
  - Balanced and Nonlinear Subspace

Recap: Nearest Neighbor Matching

- For a treated unit *i*, nearest neighbor matching (NNM) finds its nearest neighbor in control group in terms of covariates.
- NNM usually uses metrics such as Euclidean distance and Mahalanobis distance.
- NNM has difficulty in dealing with a large number of covariates. Also, bias of NNM increases with with the dimensionality of data at a rate $O(N^{-1/d})$ *[Abadie and Imbens, 2006]*

A. Abadie and G. Imbens. "Large sample properties of matching estimators for average treatment effects." Econometrica 74.1 (2006): 235-267.

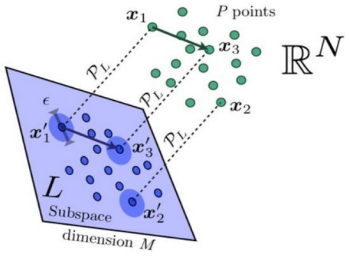
NNM with Random Subspaces

- **Motivation**
  - Dimension reduction, to soften the dependence of bias to dimension
  - Linear projection, to deal with 'big data'
- *Johnson-Lindenstrauss (JL) Lemma*

  Project data to a randomly generated subspace while preserving original distances between points *[Johnson and Lindenstrauss, 1984]*

Johnson-Lindenstrauss (JL) lemma. For any $0 < \epsilon < 1/2$ and $x_1, \cdots, x_N \in \mathbb{R}^d$, there exists a mapping $f : \mathbb{R}^d \to \mathbb{R}^k$, with $k = O(\epsilon^{-2} \log N)$, such that

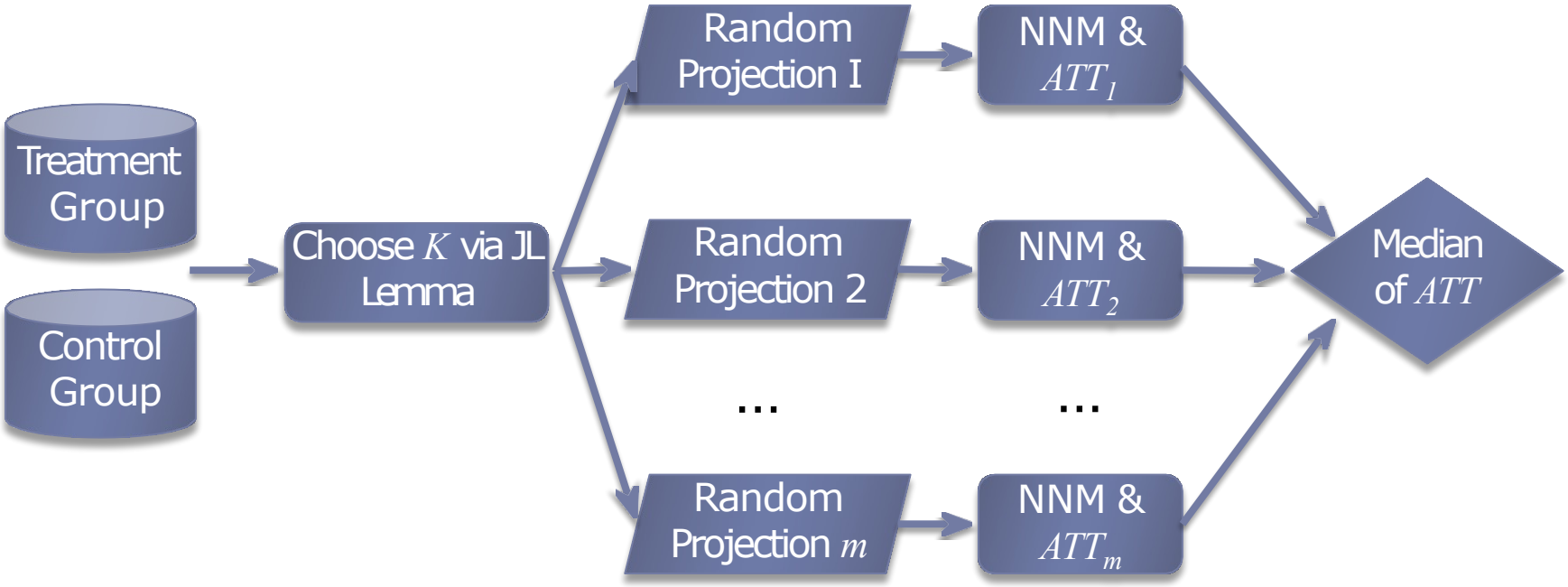
$$\forall i, j \ (1-\epsilon)\|x_i - x_j\|^2 \leq \|f(x_i) - f(x_j)\|^2 \leq (1+\epsilon)\|x_i - x_j\|^2.$$



S. Li, N. Vlassis, J. Kawale, Y. Fu. "Matching via Dimensionality Reduction for Estimation of Treatment Effects in Digital Marketing Campaigns". *IJCAI 2016*

NNM with Random Subspaces

- **Randomized Nearest Neighbor Matching (RNNM)**



Informative Subspace Learning

- Hilbert-Schmidt Independence Criterion (HSIC) based NNM
- HSIC-NNM learns two linear projections for control outcome estimation task and treated outcome estimation task separately
- It maximizes nonlinear dependency between the projected subspace and the outcome by

$$M_w = \max_{M_w} \mathrm{HSIC}(\mathbf{X}_w M_w, Y_w^F) - \mathcal{R}(M_w)$$

where X_w M_w is the transformed subspace, Y_w^F is the observed control/ treated outcome, and R denotes the regularization term

Y. Chang, J. Dy. "Informative Subspace Learning for Counterfactual Inference". *AAAI 2017*.

Nonlinear and Balanced Subspace Learning

- **Challenges**
  - Bias increases with the dimension of data
  - Complex and unbalanced distributions of high-dimensional covariates
- **Solution**
  - Encourage units with the same outcome prediction to have similar representations after transformation
  - Minimize the Maximum mean discrepancy (MMD) criterion between transformed control and treatment groups to learn balanced representations

S. Li, and Y. Fu. "Matching on balanced nonlinear representations for treatment effects estimation". *NIPS 2017*.

Summary of Subspace Learning for Causal Inference

- (+) Most methods are highly efficient owing to their closed-form solutions
- (-) Subspace learning methods usually have strong assumptions on underlying data distributions
- (-) They are usually combined with Matching estimators, but are not capable of estimating counterfactuals directly

Outline

- Regression Adjustment
  - Conditional outcome modeling
  - Grouped conditional outcome modeling
  - X-learner
- Re-weighting methods (continue)
  - CBPS
  - Trimming
  - D2VD
- Subspace Learning
- Difference-in-Difference

Difference-in-difference

- When time is involved!
- Example: After a new policy is enacted, how to estimate the effects of the policy?

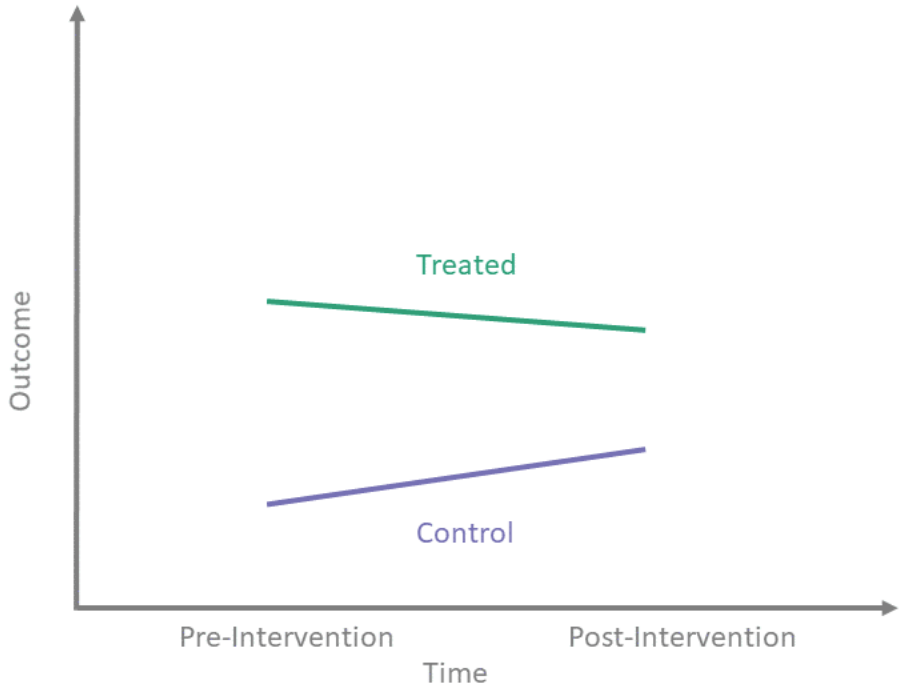
Difference-in-difference

- Average treatment effects on the treated (ATT):

$$E[Y(1) - Y(0)|T = 1]$$

Difference-in-difference

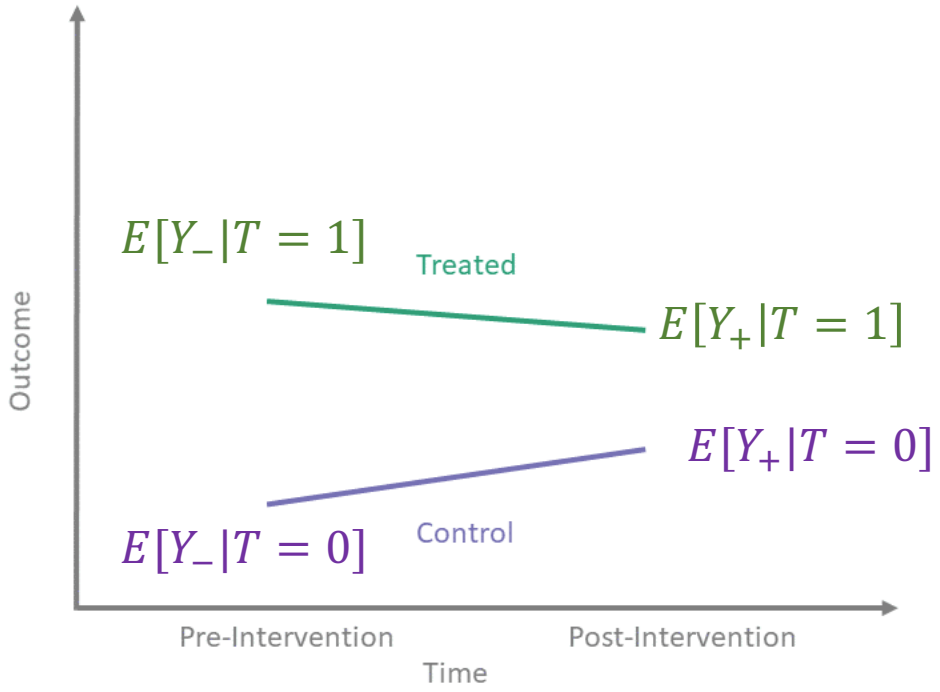
- Key: compare changes in outcomes over time between two groups



https://diff.healthpolicydatascience.org/

Difference-in-difference

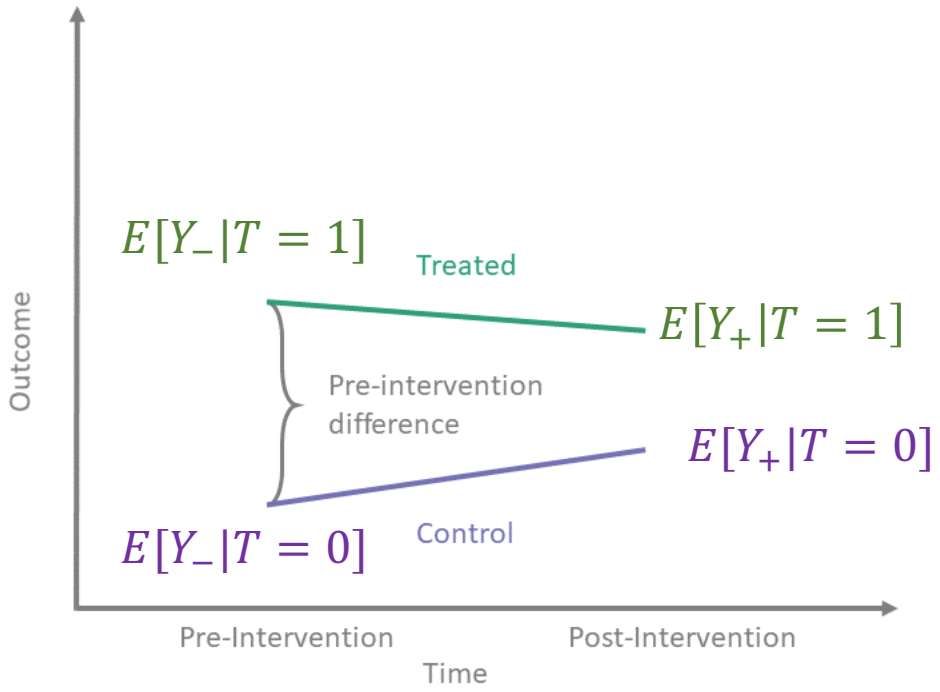
- Key: compare changes in outcomes over time between two groups



https://diff.healthpolicydatascience.org/

Difference-in-difference

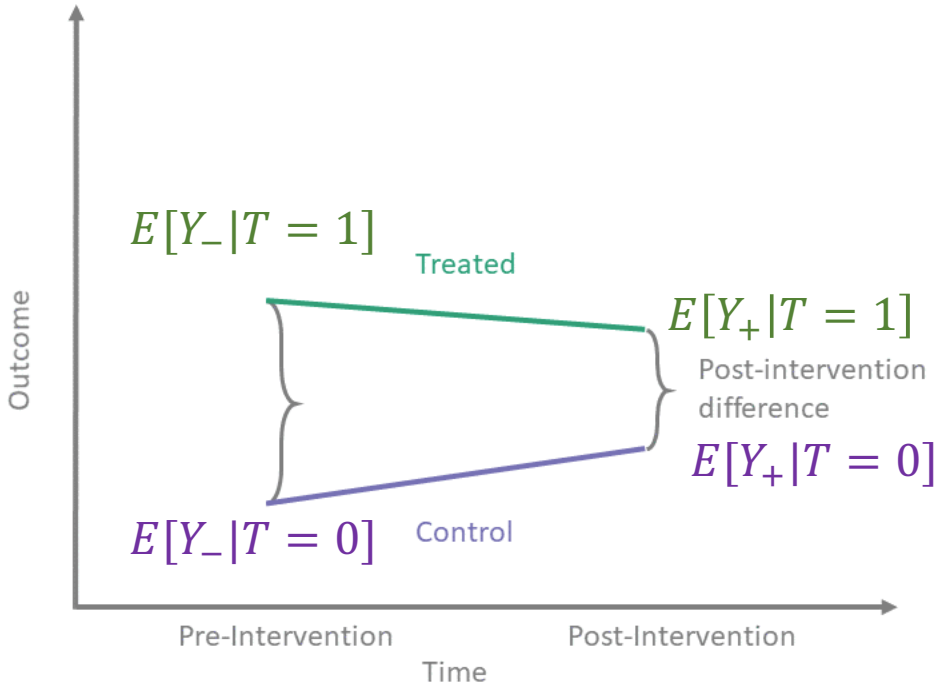
- Key: compare changes in outcomes over time between two groups



https://diff.healthpolicydatascience.org/

Difference-in-difference

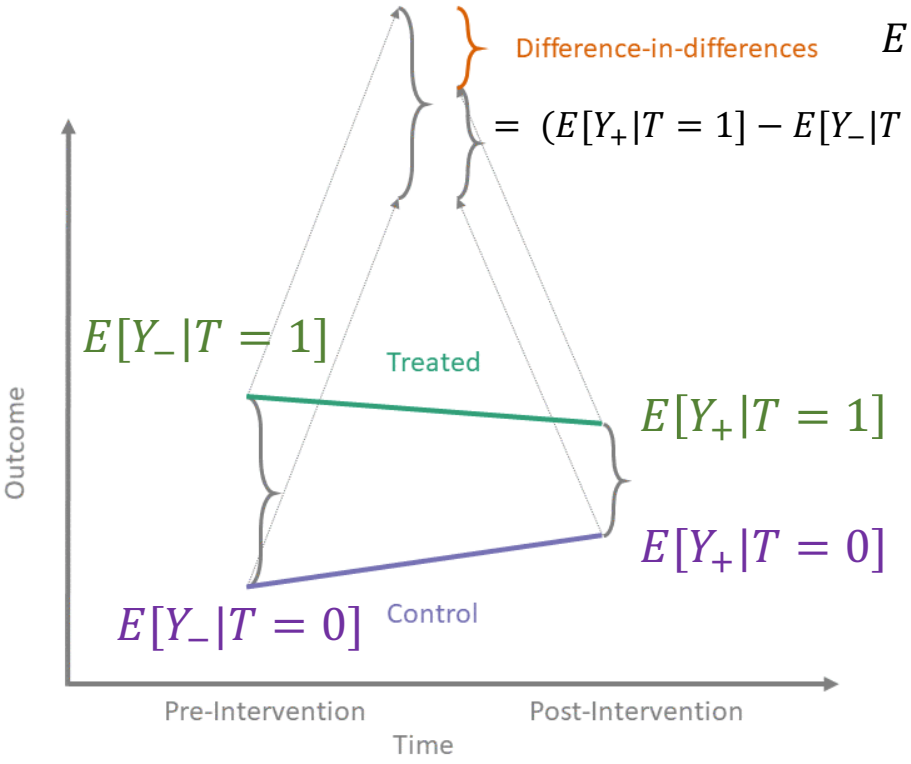
- Key: compare changes in outcomes over time between two groups



https://diff.healthpolicydatascience.org/

Difference-in-difference

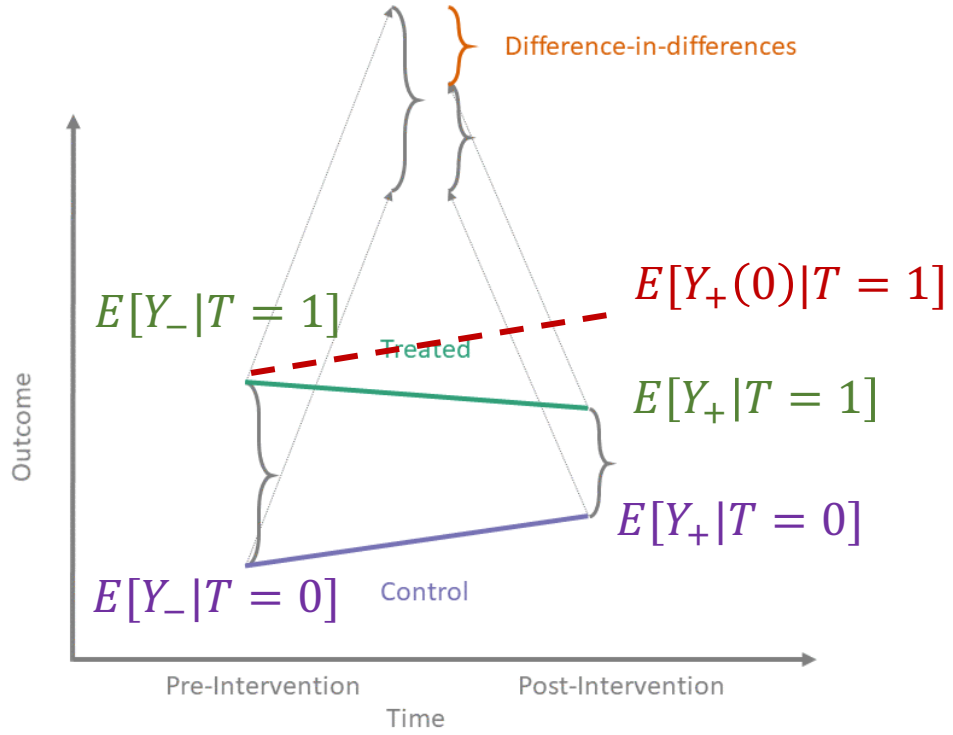
- Key: compare changes in outcomes over time between two groups



https://diff.healthpolicydatascience.org/

Difference-in-difference

- Key: compare changes in outcomes over time between two groups



https://diff.healthpolicydatascience.org/

$$E[Y_+(1) - Y_+(0) \mid T=1] = (E[Y_+|T=1] - E[Y_-|T=1]) - (E[Y_+|T=0] - E[Y_-|T=0])$$

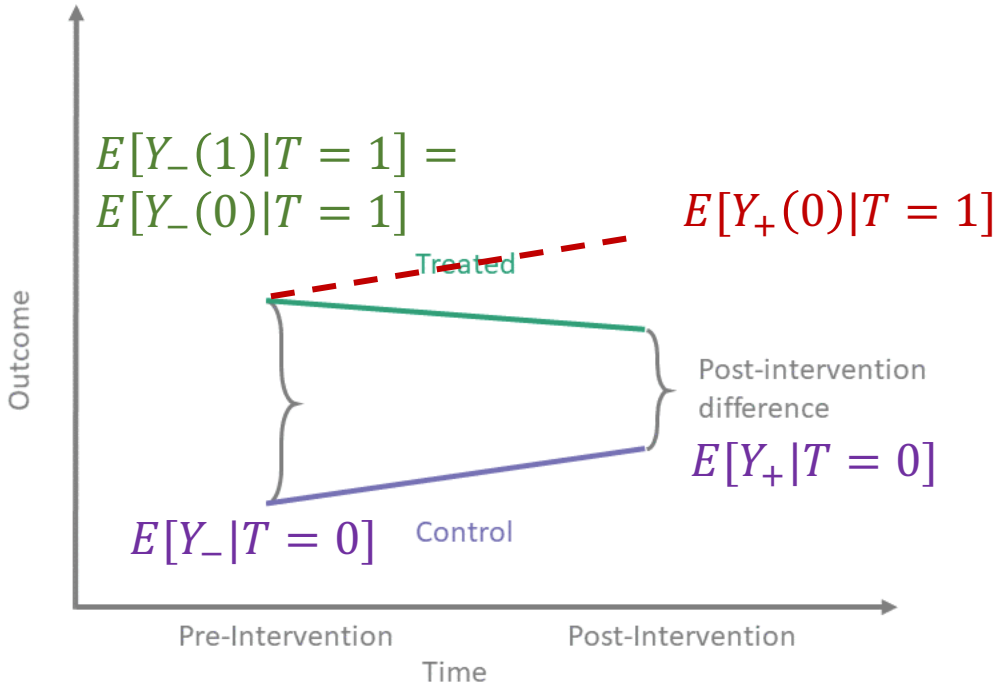
Assumptions

- Consistency assumption extended to time
  - For any timestep *, $T = t \implies Y_*(t) = Y$
  - Therefore, we have
    - $\mathrm{E}[Y_*(1)|T = 1] = E[Y|T = 1]$
    - $\mathrm{E}[Y_*(0)|T = 0] = E[Y|T = 0]$

Assumptions

- No pretreatment effect assumption

$$E[Y_{-}(1) \mid T = 1] = E[Y_{-}(0) \mid T = 1]$$

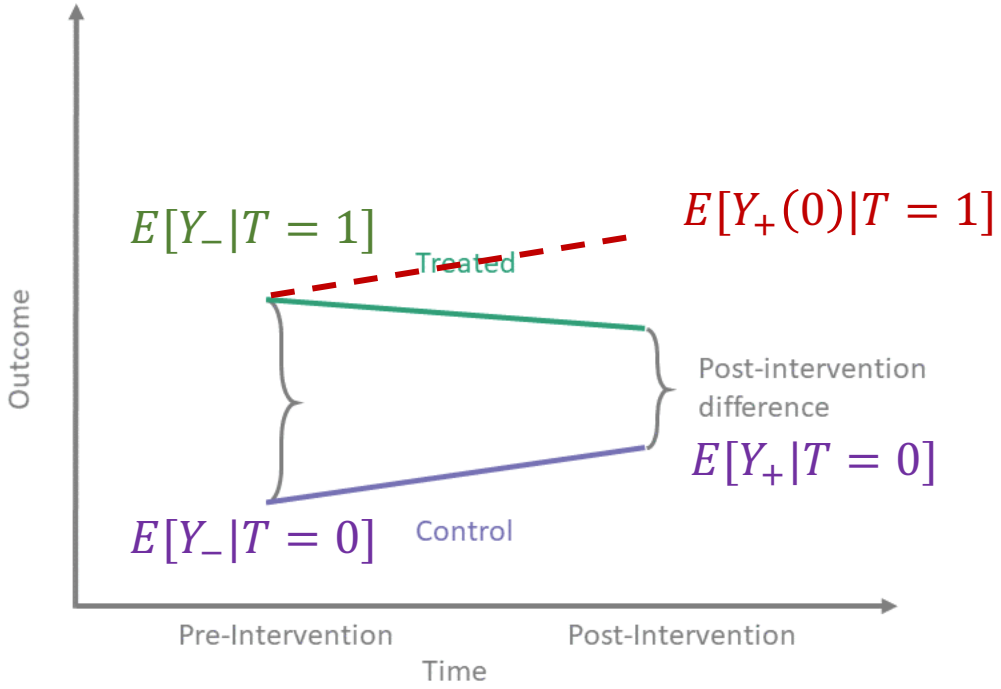


https://diff.healthpolicydatascience.org/

Assumptions

- Parallel trend assumption

$$E[Y_+(0)|T=1] - E[Y_-(0) \mid T=1] = E[Y_+(0)| T=0] - E[Y_-(0)|T=0]$$



https://diff.healthpolicydatascience.org/

Proof

$$E[Y_+(1) - Y_+(0) \mid T = 1]$$

$$= \ (E[Y_+|T = 1] - E[Y_-|T = 1]) - (E[Y_+|T = 0] - E[Y_-|T = 0])$$

Proof

$$
\begin{aligned}
& E[Y_+(1) - Y_+(0) \mid T = 1] \\
= \; & E[Y_+(1) | T = 1] - E[Y_+(0) | T = 1] \\
= \; & E[Y_+ \mid T = 1] - E[Y_+(0) | T = 1] \qquad \text{Consistency}
\end{aligned}
$$

Proof

$$
\begin{aligned}
& E[Y_+(1) - Y_+(0) \mid T = 1] \\
= \; & E[Y_+(1) | T = 1] - E[Y_+(0) | T = 1] \\
= \; & E[Y_+ \mid T = 1] - E[Y_+(0) | T = 1]
\end{aligned}
$$

Consistency

Parallel trend assumption

$$E[Y_+(0) | T = 1] = E[Y_-(0) \mid T = 1] + E[Y_+(0) | T = 0] - E[Y_-(0) | T = 0]$$

Proof

$$
\begin{aligned}
& E[Y_+(1) - Y_+(0) \mid T = 1] \\
= \; & E[Y_+(1) | T = 1] - E[Y_+(0) | T = 1] \\
= \; & E[Y_+ \mid T = 1] - E[Y_+(0) | T = 1] \qquad \text{Consistency}
\end{aligned}
$$

Parallel trend assumption

$$
\begin{aligned}
E[Y_+(0) | T = 1] = \; & E[Y_-(0) \mid T = 1] + E[Y_+(0) | \, T = 0] - E[Y_-(0) | T = 0] \\
= \; & E[Y_-(0) \mid T = 1] + E[Y_+ | \, T = 0] - E[Y_- | T = 0] \qquad \text{Consistency}
\end{aligned}
$$

Proof

$$
\begin{aligned}
& E[Y_+(1) - Y_+(0) \mid T = 1] \\
= \; & E[Y_+(1) | T = 1] - E[Y_+(0) | T = 1] \\
= \; & E[Y_+ \mid T = 1] - E[Y_+(0) | T = 1] \qquad \text{Consistency}
\end{aligned}
$$

Parallel trend assumption

$$
\begin{aligned}
E[Y_+(0) | T = 1] = \; & E[Y_-(0) \mid T = 1] + E[Y_+(0) | \, T = 0] - E[Y_-(0) | T = 0] \\
= \; & E[Y_-(0) \mid T = 1] + E[Y_+ | \, T = 0] - E[Y_- | T = 0] \qquad \text{Consistency} \\
= \; & E[Y_-(1) \mid T = 1] + E[Y_+ | \, T = 0] - E[Y_- | T = 0] \qquad \text{No pretreatment effect}
\end{aligned}
$$

Proof

$$
\begin{aligned}
& E[Y_+(1) - Y_+(0) \mid T = 1] \\
= \; & E[Y_+(1) | T = 1] - E[Y_+(0) | T = 1] \\
= \; & E[Y_+ \mid T = 1] - E[Y_+(0) | T = 1] \qquad \text{Consistency}
\end{aligned}
$$

Parallel trend assumption

$$
\begin{aligned}
E[Y_+(0) | T = 1] = \; & E[Y_-(0) \mid T = 1] + E[Y_+(0) | \, T = 0] - E[Y_-(0) | T = 0] \\
& = E[Y_-(0) \mid T = 1] + E[Y_+ | \, T = 0] - E[Y_- | T = 0] \qquad \text{Consistency} \\
& = E[Y_-(1) \mid T = 1] + E[Y_+ | \, T = 0] - E[Y_- | T = 0] \qquad \text{No pretreatment effect} \\
& = E[Y_- | \, T = 1] + E[Y_+ | \, T = 0] - E[Y_- | T = 0] \qquad \text{Consistency}
\end{aligned}
$$

Proof

$$
\begin{aligned}
& E[Y_+(1) - Y_+(0) \mid T = 1] \\
= \; & E[Y_+(1) | T = 1] - E[Y_+(0) | T = 1] \\
= \; & E[Y_+ \mid T = 1] - E[Y_+(0) | T = 1] \qquad \text{Consistency}
\end{aligned}
$$

Parallel trend assumption

$$
\begin{aligned}
E[Y_+(0) | T = 1] = \; & E[Y_-(0) \mid T = 1] + E[Y_+(0) | \, T = 0] - E[Y_-(0) | T = 0] \\
& = E[Y_-(0) \mid T = 1] + E[Y_+ | \, T = 0] - E[Y_- | T = 0] \qquad \text{Consistency} \\
& = E[Y_-(1) \mid T = 1] + E[Y_+ | \, T = 0] - E[Y_- | T = 0] \qquad \text{No pretreatment effect} \\
& = E[Y_- | \, T = 1] + E[Y_+ | \, T = 0] - E[Y_- | T = 0] \qquad \text{Consistency}
\end{aligned}
$$

$$
E[Y_+(1) - Y_+(0) \mid T = 1] \; = \; (E[Y_+ | T = 1] - E[Y_- | T = 1]) - (E[Y_+ | T = 0] - E[Y_- | T = 0])
$$

Reading materials

- Künzel S R, Sekhon J S, Bickel P J, et al. Metalearners for estimating heterogeneous treatment effects using machine learning[J]. Proceedings of the national academy of sciences, 2019, 116(10): 4156-4165.
  - https://www.pnas.org/doi/full/10.1073/pnas.1804597116
- Regression Discontinuity Designs in Economics (Lee & Lemieux, 2010)
  - https://www.princeton.edu/~davidlee/wp/RDDEconomics.pdf

Reminder of upcoming tasks

- Homework 1 is out
  - Due of HW1: Midnight Fri Sep 20 (2 weeks after being posted)

- Find a group for your paper presentation/course project!
  - 1~3 people for course project
  - 1~2 people for paper presentation
  - Course proposal due: Oct 1

Thank you!
Q&A