Machine Learning for Healthcare

6.871, HST.956

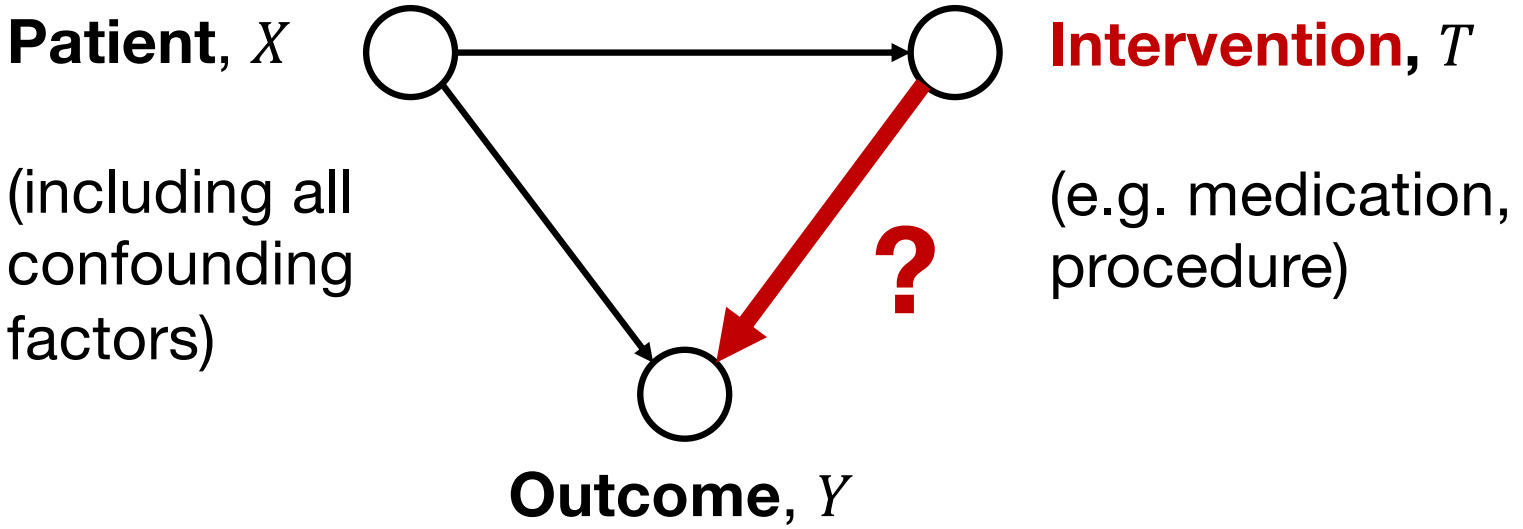
Lecture 11: Causal Inference Part 2

David Sontag

HEALTH SCIENCES
& TECHNOLOGY

Acknowledgement: some slides adapted from Uri Shalit (Technion)

Reminder: Causal inference



High dimensional

Observational data

Reminder: Potential Outcomes

- Each unit (individual) x_i has two potential outcomes:
  - $Y_0(x_i)$ is the potential outcome had the unit not been treated: "***control outcome***"
  - $Y_1(x_i)$ is the potential outcome had the unit been treated: "***treated outcome***"

- Conditional average treatment effect for unit i:
  $CATE(x_i) = \mathbb{E}_{Y_1 \sim p(Y_1|x_i)}[Y_1|x_i] - \mathbb{E}_{Y_0 \sim p(Y_0|x_i)}[Y_0|x_i]$
- Average Treatment Effect:
  $$ATE = \mathbb{E}_{x \sim p(x)}[CATE(x)]$$

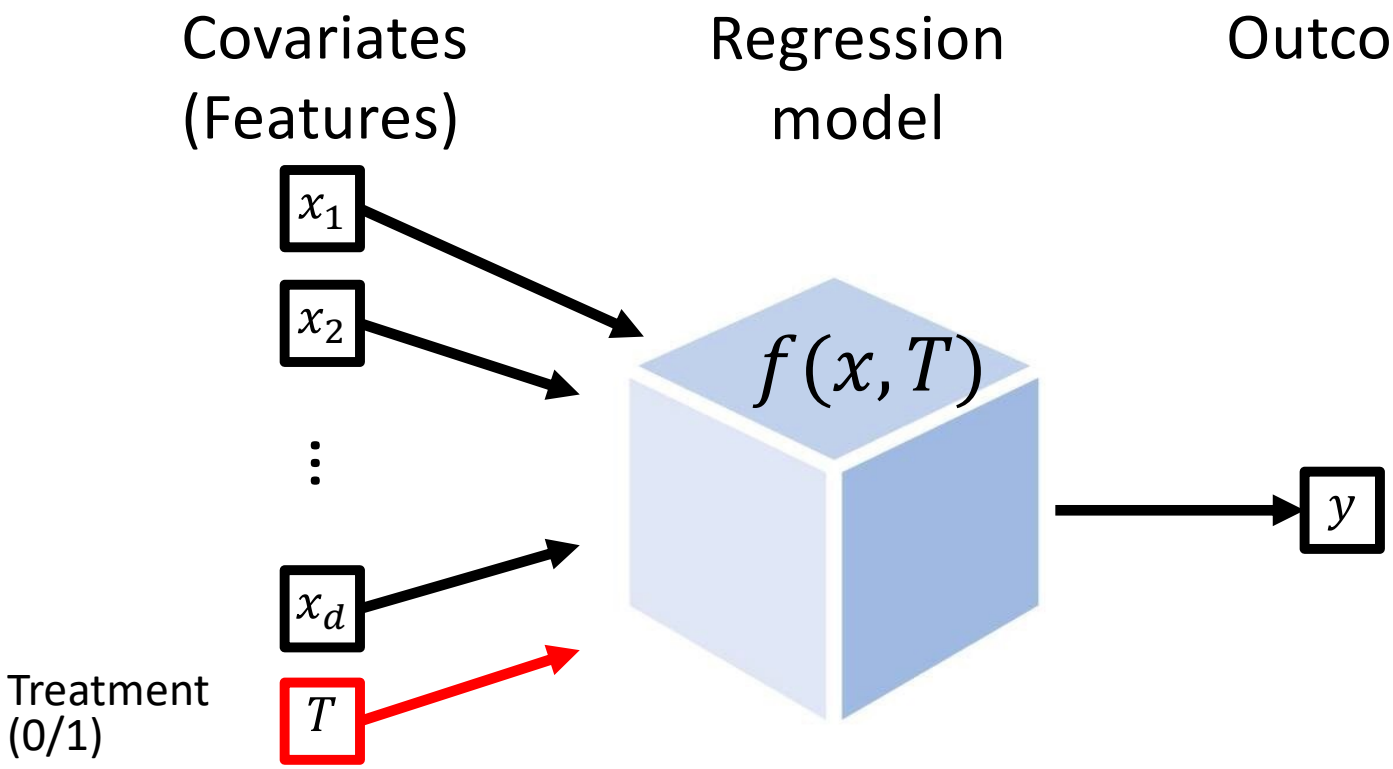
Two common approaches for counterfactual inference

Covariate adjustment

Propensity scores

Covariate adjustment (reminder)

Explicitly model the relationship between treatment, confounders, and outcome:



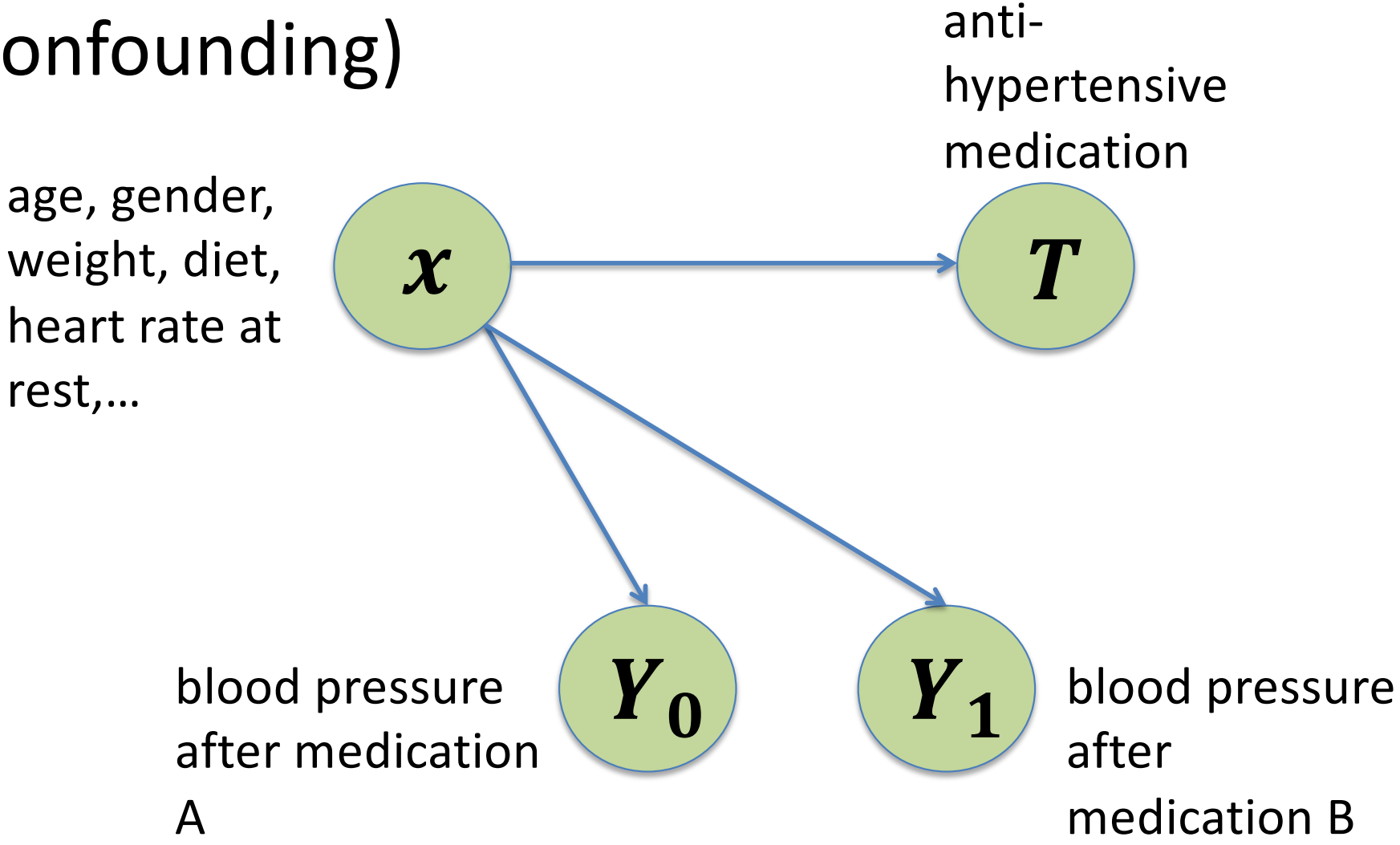
Covariate adjustment (reminder)

- Under ignorability, can use the adjustment formula:

$$ATE(x) = \mathbb{E}_{x\sim p(x)}\left[\mathbb{E}[Y_1|T=1,x] - \mathbb{E}[Y_0|T=0,x]\right]$$

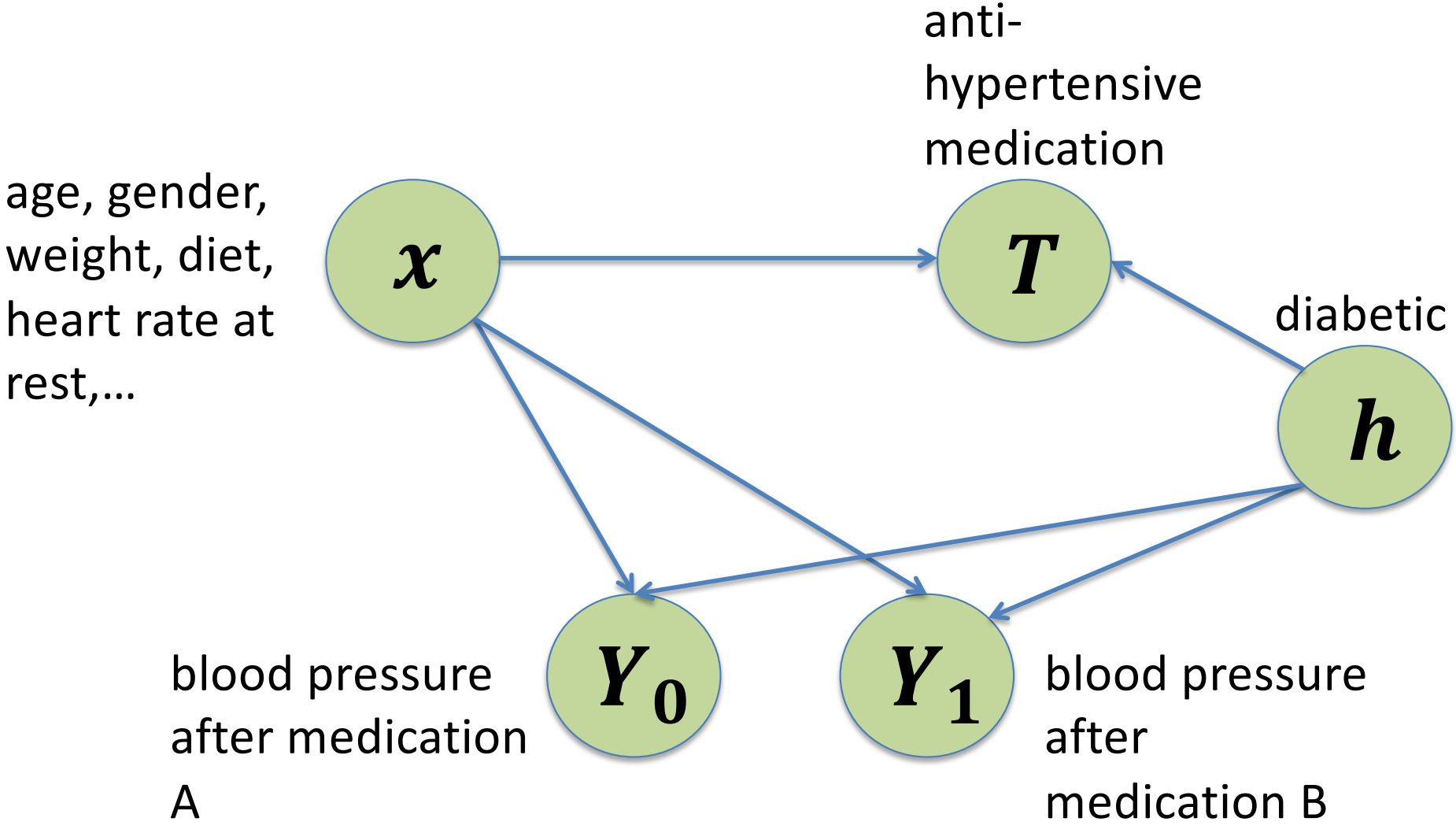
- Fit a model $f(x,t) \approx \mathbb{E}[Y_t|T=t,x]$, then: $\widehat{CATE}(x) = f(x,1) - f(x,0)$.

Ignorability (no hidden confounding)



$$(Y_0, Y_1) \perp\!\!\!\perp T \mid x$$

No Ignorability



$$(Y_0, Y_1) \not\perp\!\!\!\perp T \mid x$$

Covariate adjustment with linear models

- Assume that:

Blood pressure **age** **medication**

$$Y_t(x) = \beta x + \gamma \cdot t + \epsilon_t$$

$$\mathbb{E}[\epsilon_t] = 0$$

- Then:

$$CATE(x) := \mathbb{E}[Y_1(x) - Y_0(x)] =$$

Covariate adjustment with linear models

- Assume that:

Blood pressure **age** **medication**

$$Y_t(x) = \beta x + \gamma \cdot t + \epsilon_t$$
$$\mathbb{E}[\epsilon_t] = 0$$

- Then:

$$CATE(x) := \mathbb{E}[Y_1(x) - Y_0(x)] =$$
$$\mathbb{E}[(\cancel{\beta x} + \gamma + \epsilon_1) - (\cancel{\beta x} + \epsilon_0)] = \gamma$$

$$ATE := \mathbb{E}_{p(x)}[CATE(x)] = \gamma$$

Covariate adjustment with linear models

- Assume that:

Blood pressure **age** **medication**

$$Y_t(x) = \beta x + \gamma \cdot t + \epsilon_t$$
$$\mathbb{E}[\epsilon_t] = 0$$

$$ATE := \mathbb{E}_{p(x)}[CATE(x)] = \gamma$$

- For causal inference, need to estimate γ well, not $Y_t(x)$ **- Identification, not prediction**
- *Major difference between ML and statistics*

What happens when there is misspecification?

- True data generating process, $x \in \mathbb{R}$:

$$Y_t(x) = \beta x + \gamma \cdot t + \delta \cdot x^2$$

$$ATE = \mathbb{E}[Y_1 - Y_0] = \gamma$$

- Hypothesized model:

$$\widehat{Y}_t(x) = \hat{\beta} x + \hat{\gamma} \cdot t$$

$$\hat{\gamma} = \gamma + \delta \frac{\mathbb{E}[xt]\mathbb{E}[x^2] - \mathbb{E}[t^2]\mathbb{E}[x^2 t]}{\mathbb{E}[xt]^2 - \mathbb{E}[x^2]\mathbb{E}[t^2]}$$

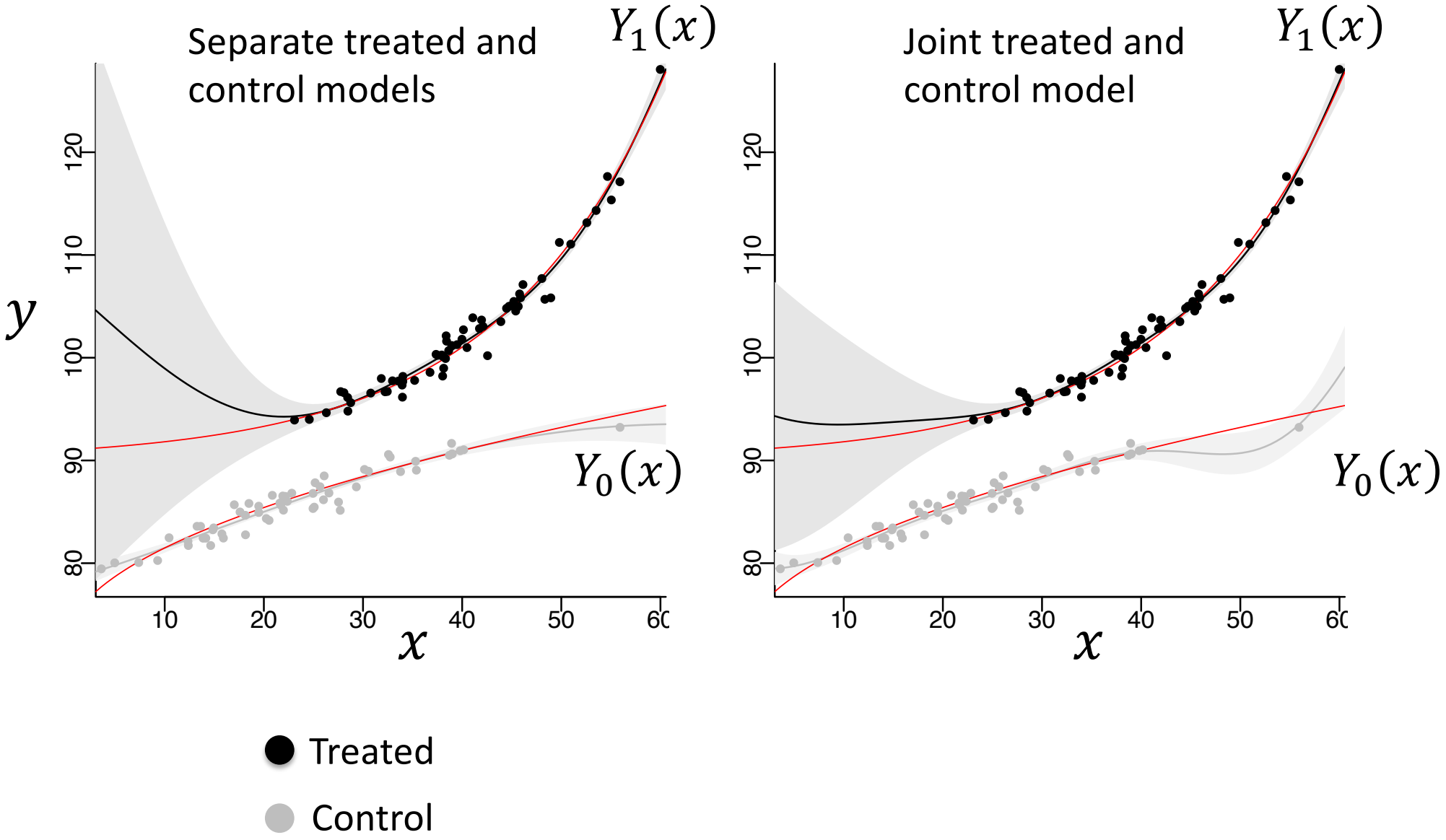
Depending on δ, can be made to be arbitrarily large or small!

Covariate adjustment with non-linear models

- Random forests and Bayesian trees
  Hill (2011), Athey & Imbens (2015), Wager & Athey (2015)
- Gaussian processes
  Hoyer et al. (2009), Zigler et al. (2012), Alaa & van der Schaar (2017)
- Neural networks
  Beck et al. (2000), Johansson et al. (2016), Shalit et al. (2016), Lopez-Paz et al. (2016)

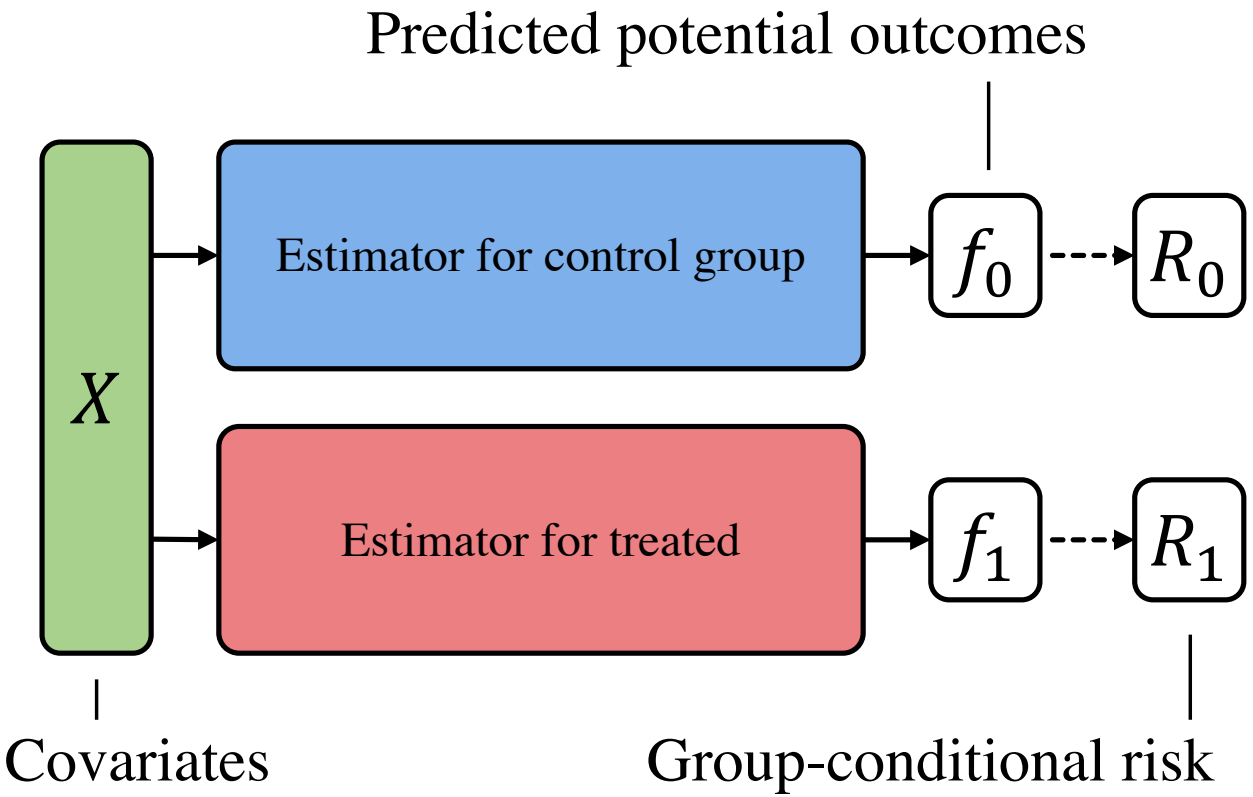
Called *nonparametric* estimators, since they do not make assumptions about form of $\mathbb{E}[Y|X,T]$ and, given enough data, could fit any function

Example: Gaussian processes

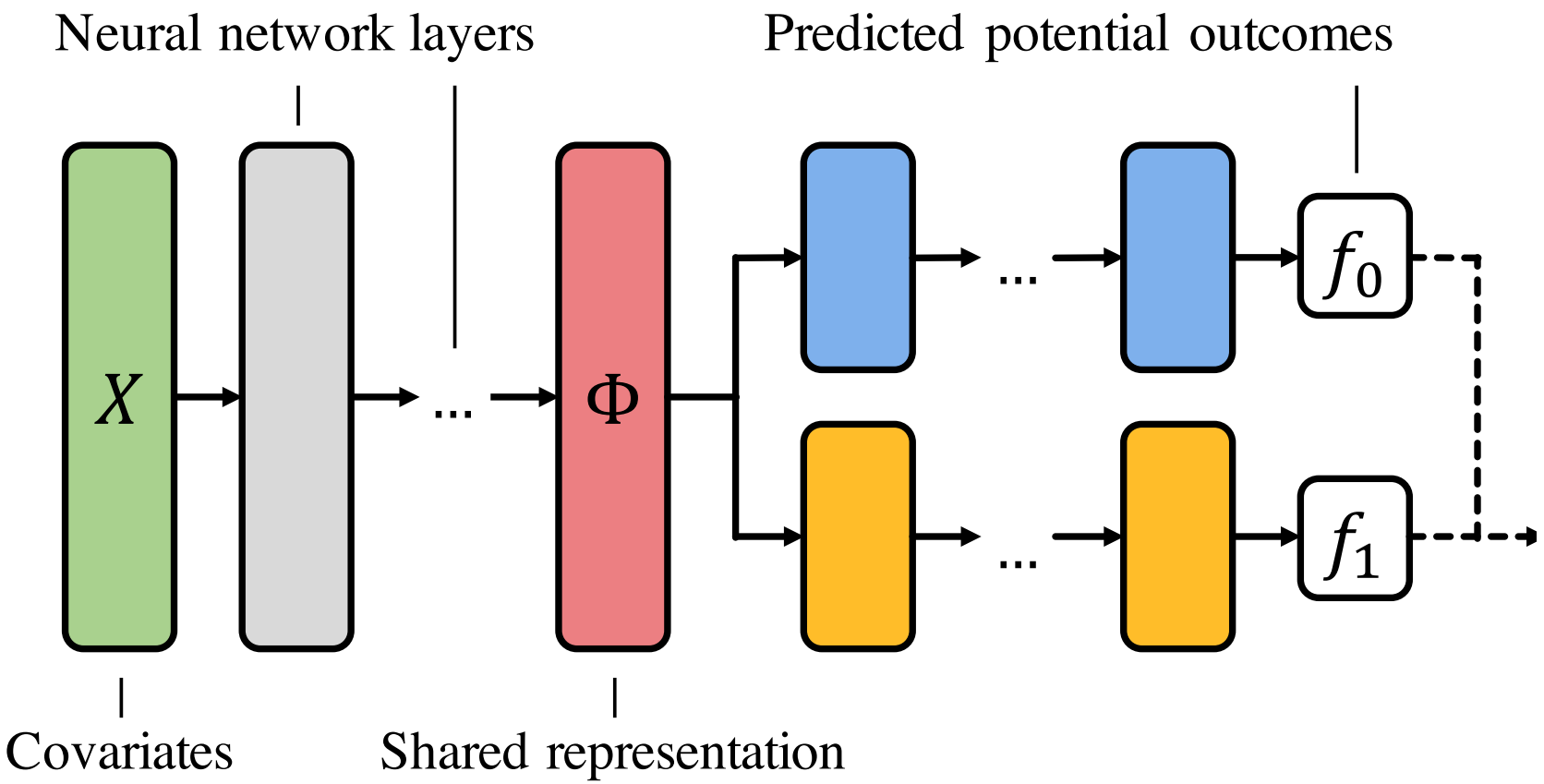


Figures: Vincent Dorie & Jennifer Hill

Example: Neural networks



Example: Neural networks



Shalit, Johansson, Sontag. *Estimating Individual Treatment Effect: Generalization Bounds and Algorithms*. ICML, 2017

Necessary assumption for nonparametric estimation – common support

Y_0, Y_1: potential outcomes for control and treated

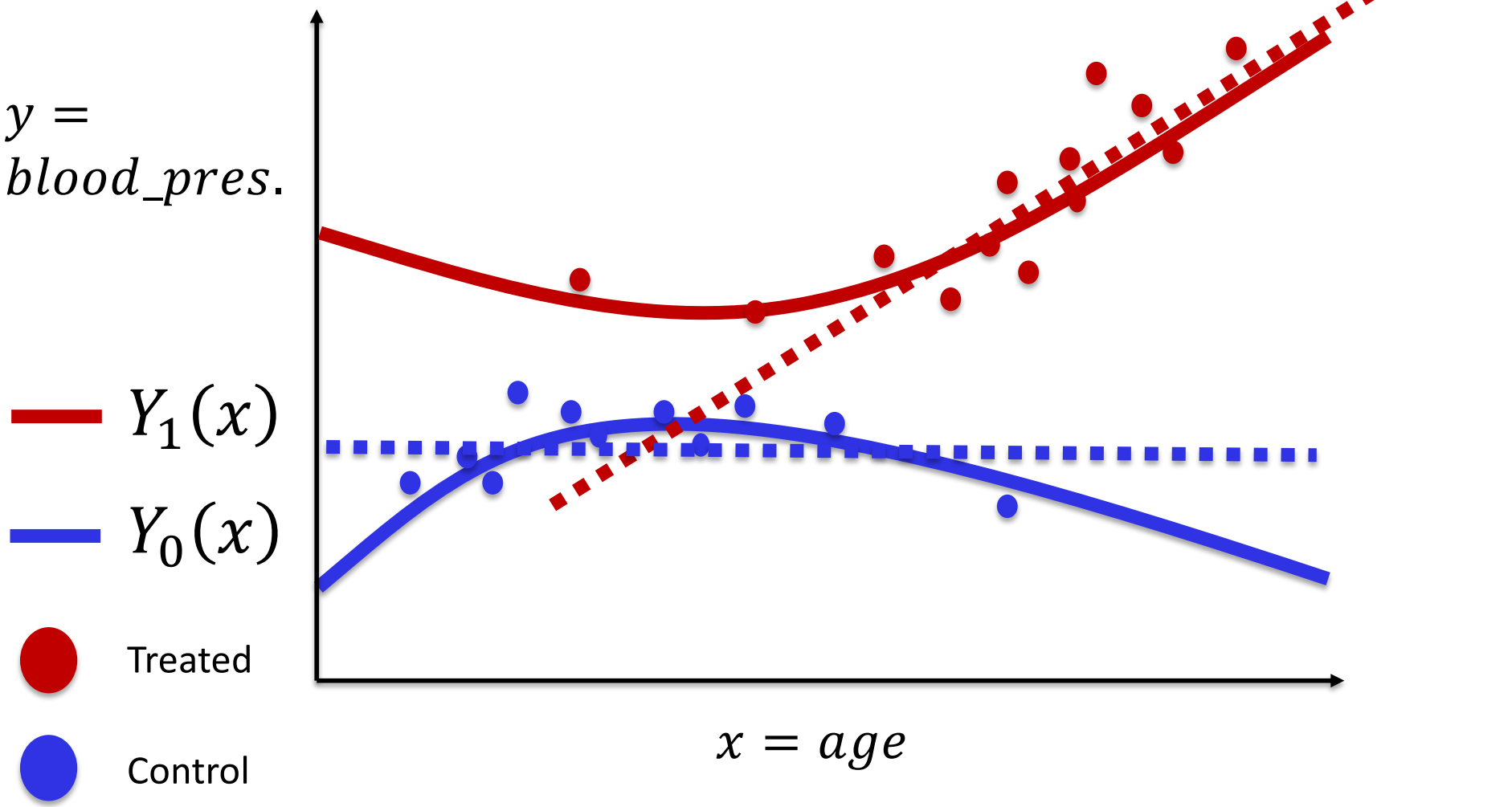
x: unit covariates (features)

T: treatment assignment

We assume:

$$p(T = t|X = x) > 0 \; \forall t, x$$

Example of how (nonparametric) covariate adjustment fails when there is no common support (overlap)



Matching

- Find each unit's long-lost counterfactual identical twin, check up on his outcome

Matching

- Find each person's long-lost counterfactual identical twin, check up on his outcome

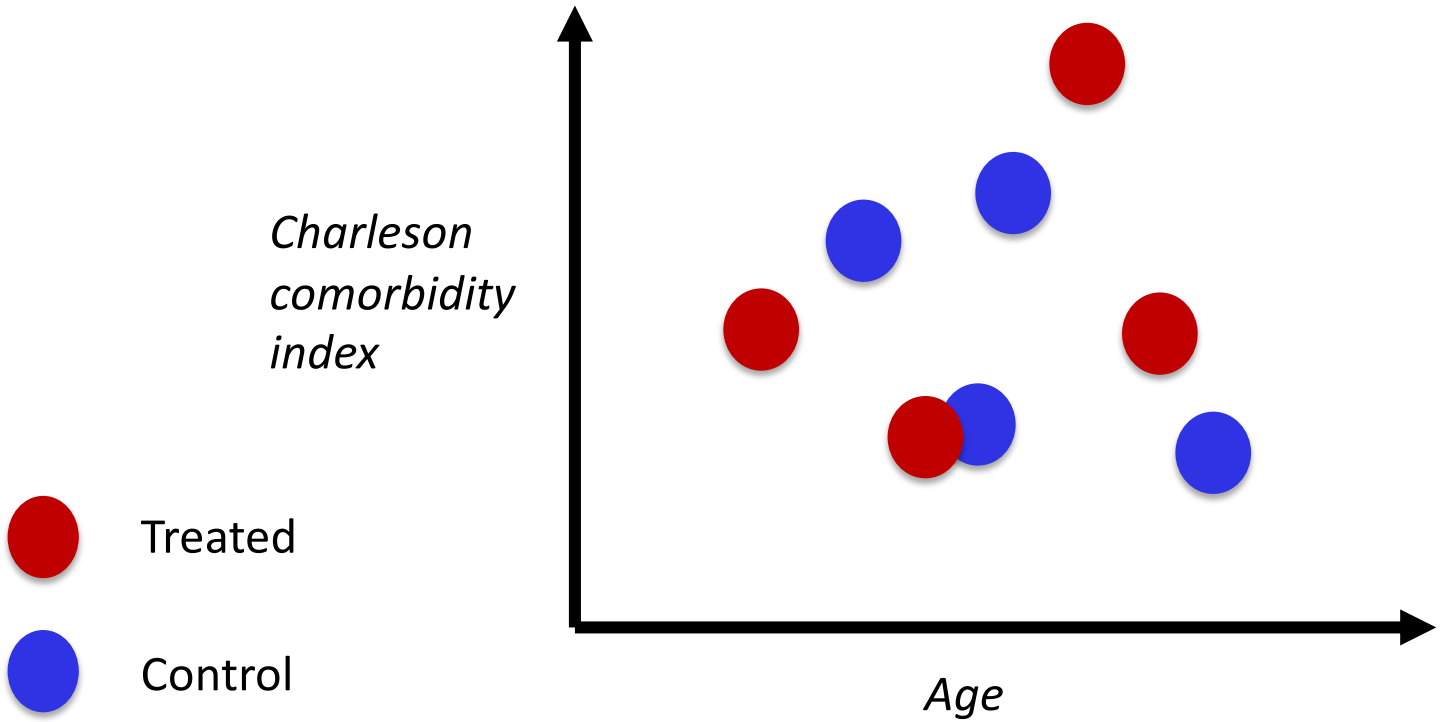
Obama, had he gone to law school

Obama, had he gone to business school

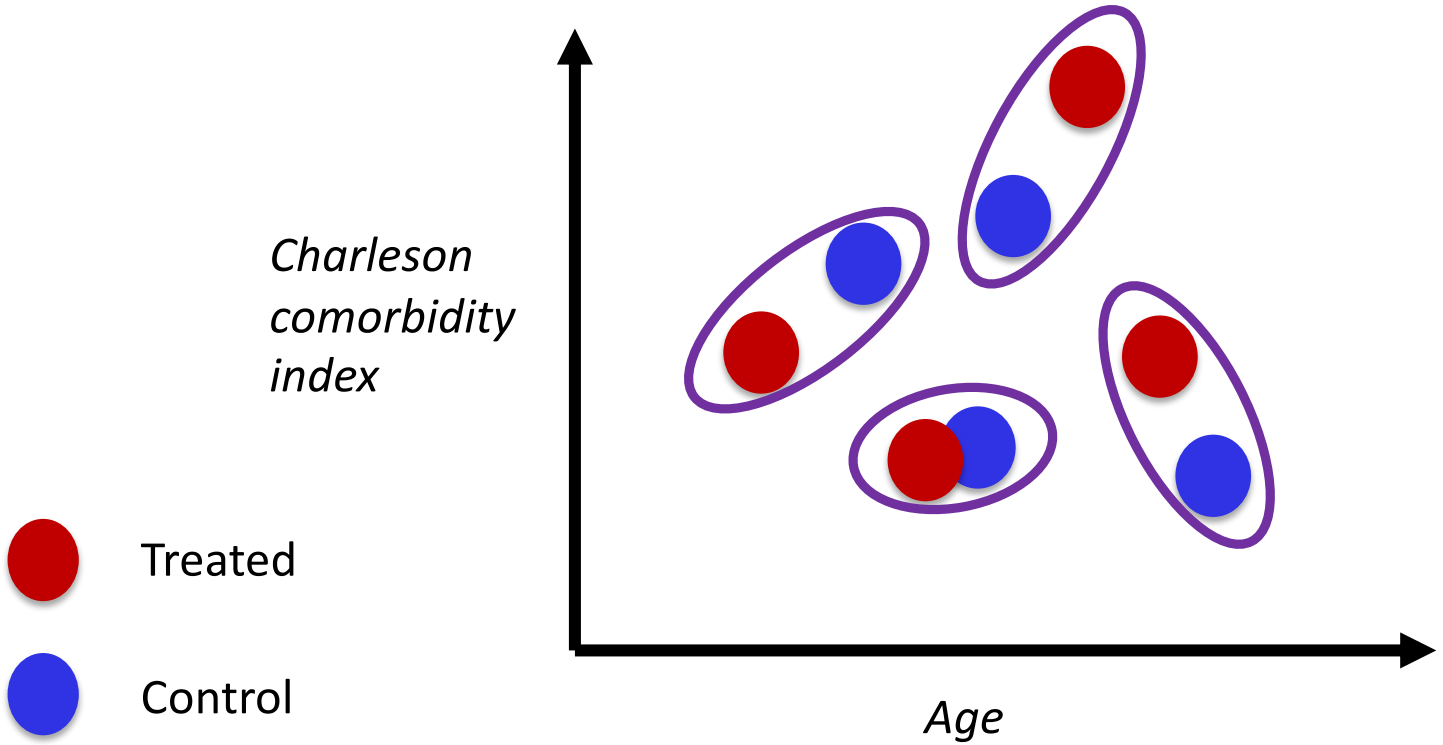
Matching

- Find each person's long-lost counterfactual identical twin, check up on his outcome
- Used for estimating both ATE and CATE

Match to nearest neighbor from opposite group



Match to nearest neighbor from opposite group



1-NN Matching

- Let $d(\cdot,\cdot)$ be a metric between x's
- For each i, define $j(i) = \underset{j \text{ s.t. } t_j \neq t_i}{\operatorname{argmin}} \; d(x_j, x_i)$

  $j(i)$ is the nearest counterfactual neighbor of i
- $t_i = 1$, unit i is treated:

$$\widehat{CATE}(x_i) = y_i - y_{j(i)}$$

- $t_i = 0$, unit i is control:

$$\widehat{CATE}(x_i) = y_{j(i)} - y_i$$

1-NN Matching

- Let $d(\cdot,\cdot)$ be a metric between x's
- For each i, define $j(i) = \underset{j\ s.t.\ t_j \neq t_i}{\operatorname{argmin}}\ d(x_j, x_i)$

  $j(i)$ is the nearest counterfactual neighbor of i

- $\widehat{CATE}(x_i) = (2t_i - 1)(y_i - y_{j(i)})$
- $\widehat{ATE} = \frac{1}{n}\sum_{i=1}^{n} \widehat{CATE}(x_i)$

Matching

- Interpretable, especially in small-sample regime
- Nonparametric
- Heavily reliant on the underlying metric
- Could be misled by features which don’t affect the outcome

Covariate adjustment and matching

- Matching is equivalent to covariate adjustment with two 1-nearest neighbor classifiers:
  $\hat{Y}_1(x) = y_{NN_1(x)}$, $\hat{Y}_0(x) = y_{NN_0(x)}$
  where $y_{NN_t(x)}$ is the nearest-neighbor of x among units with treatment assignment
  $$t = 0,1$$

- 1-NN matching is in general inconsistent, though only with small bias (Imbens 2004)

Two common approaches for counterfactual inference

Covariate adjustment

Propensity scores

Propensity scores

- Tool for estimating ATE
- Imagine that we had data from a randomized control trial (RCT). Then we could simply estimate the ATE using:

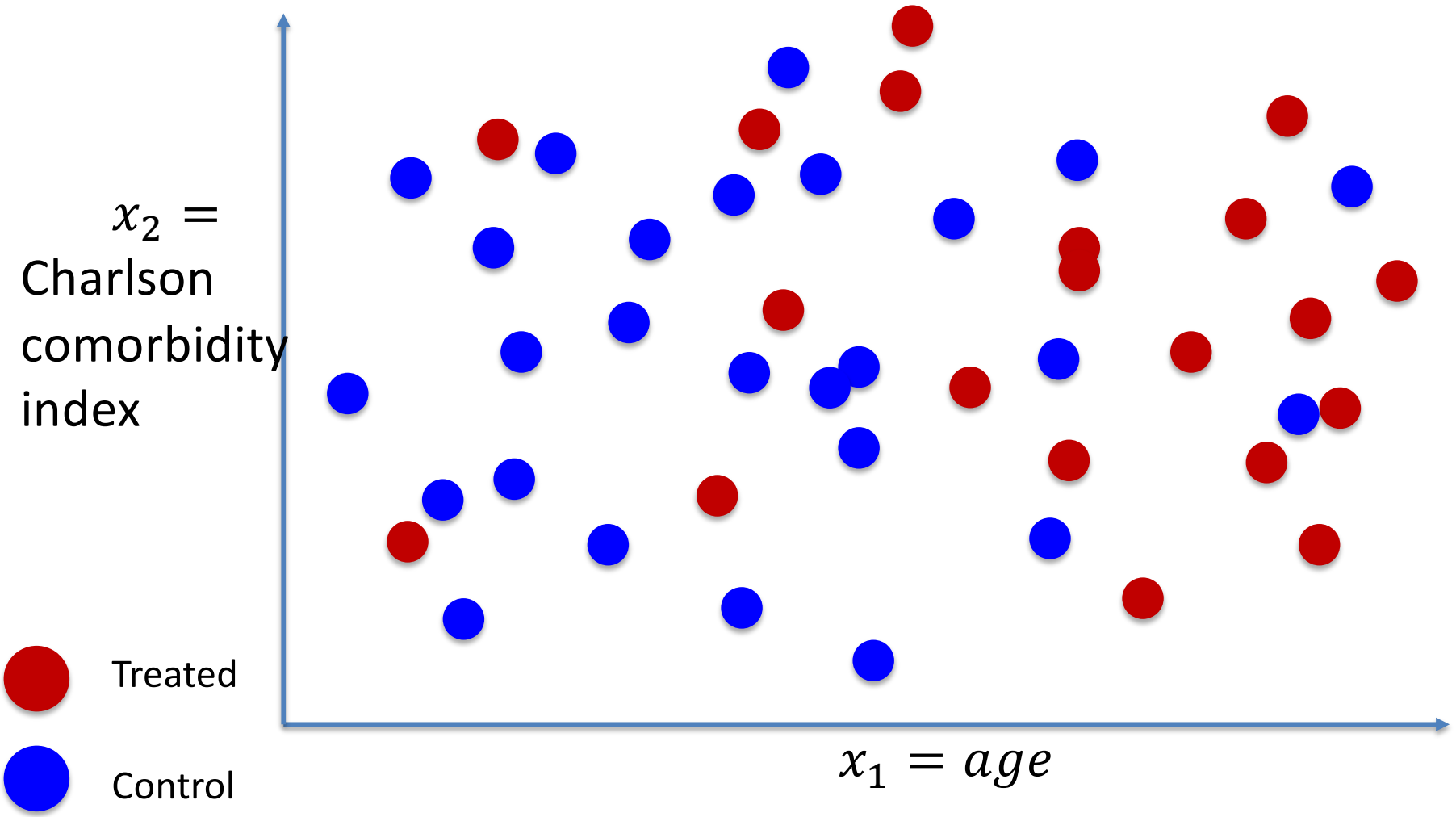
$$\frac{1}{n_1}\sum_{i\ s.t.T_i=1} Y_i - \frac{1}{n_0}\sum_{i\ s.t.T_i=0} Y_i$$

- Basic idea: turn observational study into a pseudo-randomized trial by re-weighting samples

Inverse propensity score re-weighting

$$p(x|t=0) \neq p(x|t=1)$$

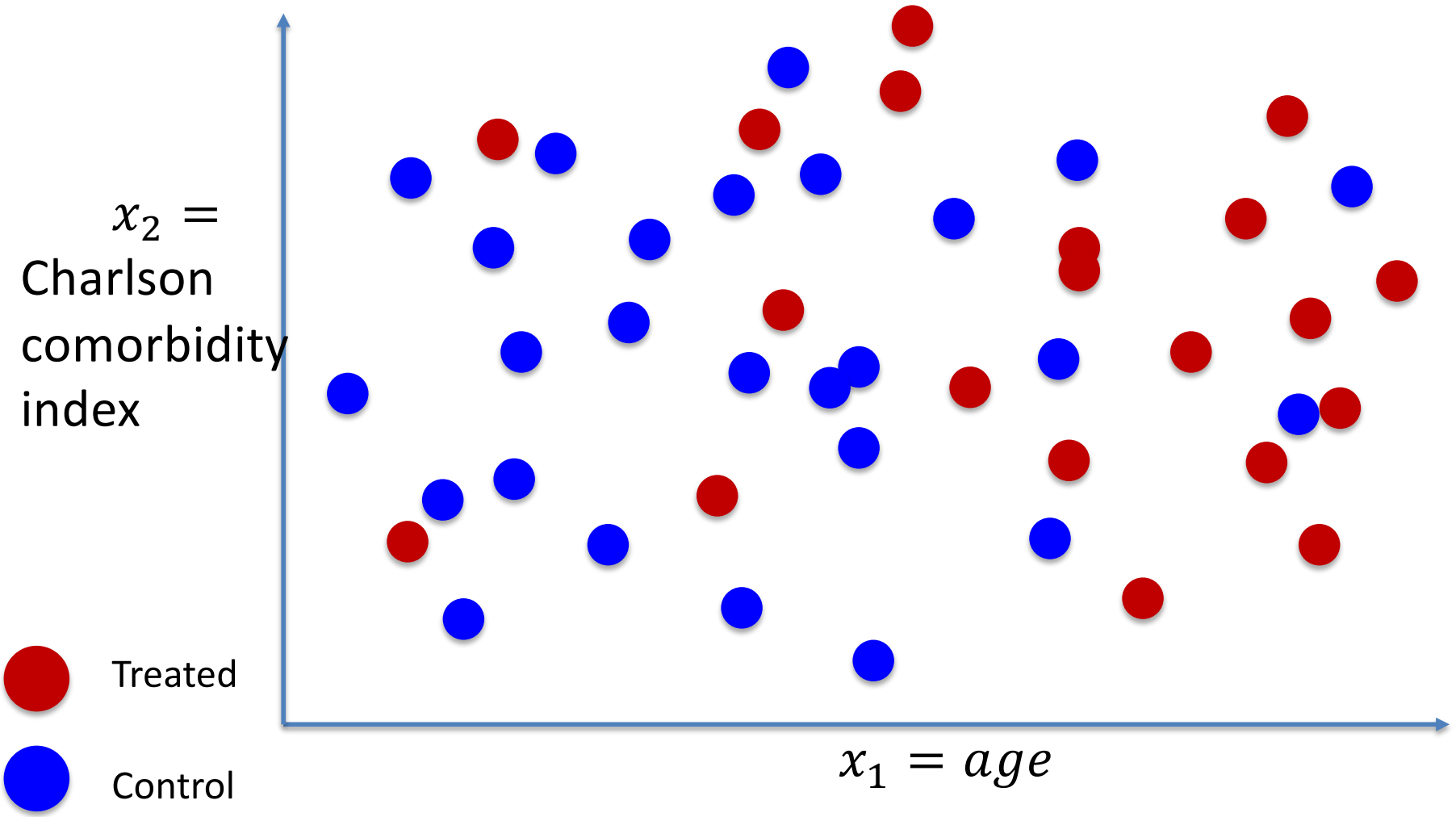
control *treated*



Inverse propensity score re-weighting

$$p(x|t=0) \cdot w_0(x) \approx p(x|t=1) \cdot w_1(x)$$

reweighted control *reweighted treated*



Propensity score

- Propensity score: $p(T = 1|x)$,
  using machine learning tools, e.g. logistic regression
- Samples re-weighted by the inverse propensity score of the treatment they received

Propensity scores – algorithm

Inverse probability of treatment weighted estimator

How to calculate ATE with propensity score for sample $(x_1, t_1, y_1), \ldots, (x_n, t_n, y_n)$

1. Use any ML method to estimate $\hat{p}(T = t|x)$

2. $$\hat{ATE} = \frac{1}{n} \sum_{i \text{ s.t. } t_i=1} \frac{y_i}{\hat{p}(t_i = 1|x_i)} - \frac{1}{n} \sum_{i \text{ s.t. } t_i=0} \frac{y_i}{\hat{p}(t_i = 0|x_i)}$$

Propensity scores – algorithm

Inverse probability of treatment weighted estimator

How to calculate ATE with propensity score for sample $(x_1, t_1, y_1), \ldots, (x_n, t_n, y_n)$

1. Randomized trial $p(T = t|x) = 0.5$

2. $$\hat{ATE} = \frac{1}{n} \sum_{i \text{ s.t. } t_i=1} \frac{y_i}{\hat{p}(t_i = 1|x_i)} - \frac{1}{n} \sum_{i \text{ s.t. } t_i=0} \frac{y_i}{\hat{p}(t_i = 0|x_i)}$$

Propensity scores – algorithm

Inverse probability of treatment weighted estimator

How to calculate ATE with propensity score for sample $(x_1, t_1, y_1), \ldots, (x_n, t_n, y_n)$

1. Randomized trial $p(T = t|x) = 0.5$

2. $\hat{ATE} = \frac{1}{n} \sum_{i \text{ s.t. } t_i=1} \frac{y_i}{0.5} - \frac{1}{n} \sum_{i \text{ s.t. } t_i=0} \frac{y_i}{0.5} =$

Propensity scores – algorithm

Inverse probability of treatment weighted estimator

How to calculate ATE with propensity score for sample $(x_1, t_1, y_1), \ldots, (x_n, t_n, y_n)$

1. Randomized trial $p(T = t|x) = 0.5$

2. $$\hat{ATE} = \frac{1}{n} \sum_{i \text{ s.t. } t_i=1} \frac{y_i}{0.5} - \frac{1}{n} \sum_{i \text{ s.t. } t_i=0} \frac{y_i}{0.5} =$$
$$\frac{2}{n} \sum_{i \text{ s.t. } t_i=1} y_i - \frac{2}{n} \sum_{i \text{ s.t. } t_i=0} y_i$$

Propensity scores – algorithm

Inverse probability of treatment weighted estimator

How to calculate ATE with propensity score for sample $(x_1, t_1, y_1), \ldots, (x_n, t_n, y_n)$

1. Randomized trial $p = 0.5$

2. $$\hat{ATE} = \frac{1}{n} \sum_{i \text{ s.t. } t_i=1} \frac{y_i}{0.5} - \frac{1}{n} \sum_{i \text{ s.t. } t_i=0} \frac{y_i}{0.5} =$$

$$\frac{2}{n} \sum_{i \text{ s.t. } t_i=1} y_i - \frac{2}{n} \sum_{i \text{ s.t. } t_i=0} y_i$$

Sum over $\sim \frac{n}{2}$ terms

Propensity scores - derivation

- How do we derive this estimator?

$$\hat{ATE} = \frac{1}{n} \sum_{i \text{ s.t. } t_i=1} \frac{y_i}{\hat{p}(t_i = 1|x_i)} - \frac{1}{n} \sum_{i \text{ s.t. } t_i=0} \frac{y_i}{\hat{p}(t_i = 0|x_i)}$$

- Recall definition of average treatment effect:

$$ATE = \mathbb{E}_{x \sim p(x)}[Y_1(x)] - \mathbb{E}_{x \sim p(x)}[Y_0(x)]$$

- Naively, using observed data we can estimate

$$\mathbb{E}_{x \sim p(x|T=1)}[Y_1(x)] \quad \& \quad \mathbb{E}_{x \sim p(x|T=0)}[Y_0(x)]$$

Propensity scores - derivation

- We want: $\mathbb{E}_{x \sim p(x)}[Y_1(x)]$
- We know that:

$$p(x|T=1) \cdot \frac{p(T=1)}{p(T=1|x)} = p(x)$$

- Thus:

$$\mathbb{E}_{x \sim p(x|T=1)} \left[\frac{p(T=1)}{p(T=1 \mid x)} Y_1(x) \right] = \mathbb{E}_{x \sim p(x)}[Y_1(x)]$$

- We can approximate this empirically as:

$$\frac{1}{n_1} \sum_{i \text{ s.t.} t_i=1} \left[\frac{n_1/n}{\hat{p}(t_i=1 \mid x_i)} y_i \right] = \frac{1}{n} \sum_{i \text{ s.t.} t_i=1} \frac{y_i}{\hat{p}(t_i=1 \mid x_i)}$$

(similarly for $t_i=0$)

Problems with inverse propensity weighting (IPW)

- Need to estimate propensity score (problem in all propensity score methods)
- If there's not much overlap, propensity scores become non-informative and easily mis-calibrated
- Weighting by inverse can create large variance and large errors for small propensity scores
  - Exacerbated when more than two treatments

Summary

- Two approaches to use machine learning for causal inference
  - Predict outcome given features and treatment, then use resulting model to impute counterfactuals (*covariate adjustment*)
  - Predict treatment using features (*propensity score*), then use to reweight outcome or stratify the data
- Consistency of estimates depend on:
  - Causal graph being correct (i.e., no unobserved confounding)
  - Identifiability of causal effect (i.e., overlap)
  - Nonparametric regression is used (or correctly specified model)

References

- Recent work from ML community: https://sites.google.com/view/nips2018causallearning/ and http://tripods.cis.cornell.edu/neurips19_causalml/
- Recent book on causal inference by Miguel Hernan and Jamie Robins: https://www.hsph.harvard.edu/miguel-hernan/causal-inference-book/
  Recent book on causal inference by Jonas Peters, Dominik Janzing and Bernhard Schölkopf: https://mitpress.mit.edu/books/elements-causal-inference
  (download PDF for free on left: “Open Access Title”)
- Examples of recent papers in this research field:
  https://arxiv.org/abs/1906.02120
  https://arxiv.org/abs/1705.08821
  https://arxiv.org/abs/1510.04342
  https://arxiv.org/abs/1810.02894