

Machine Learning for Healthcare

HST.956, 6.S897

Lecture 14: Causal Inference Part 2

David Sontag



Acknowledgement: adapted from slides by Uri Shalit (Technion)

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:

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

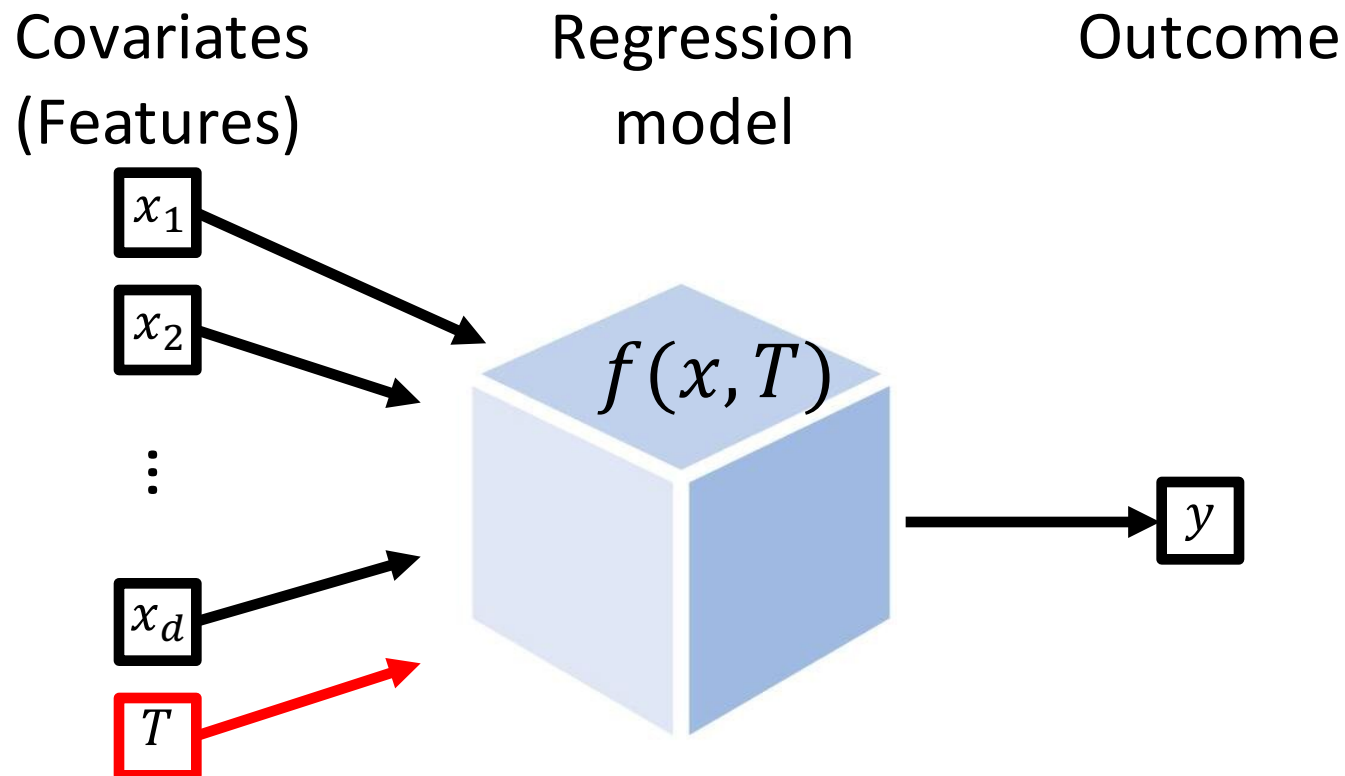
Two common approaches for counterfactual inference

Covariate adjustment

Propensity scores

Covariate adjustment (reminder)

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



Covariate adjustment (reminder)

- Under ignorability,

$$CATE(x) =$$

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

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

$$\widehat{CATE}(x_i) = f(x_i, 1) - f(x_i, 0).$$

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:

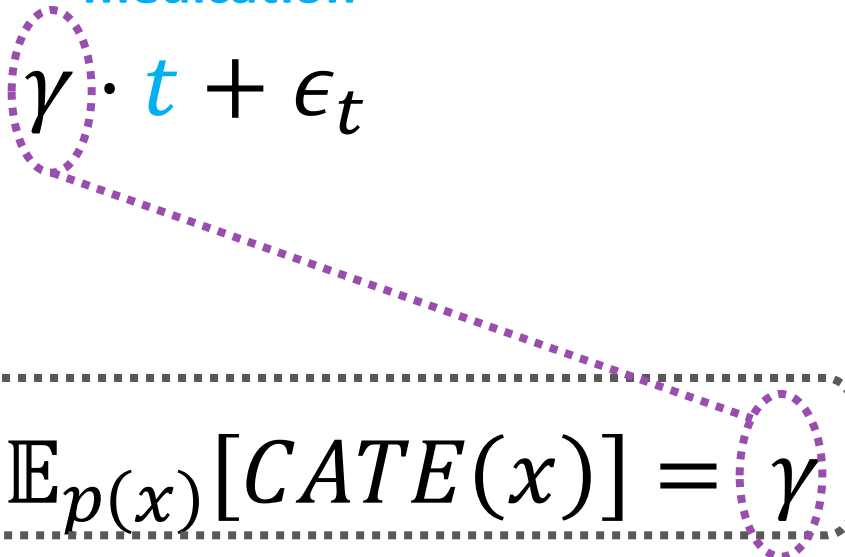
$$\begin{aligned} CATE(x) &:= \mathbb{E}[Y_1(x) - Y_0(x)] = \\ \mathbb{E}[(\cancel{\beta x} + \gamma + \epsilon_1) - (\cancel{\beta x} + \epsilon_0)] &= \gamma \end{aligned}$$

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

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 if true model is not linear?

- 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:

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

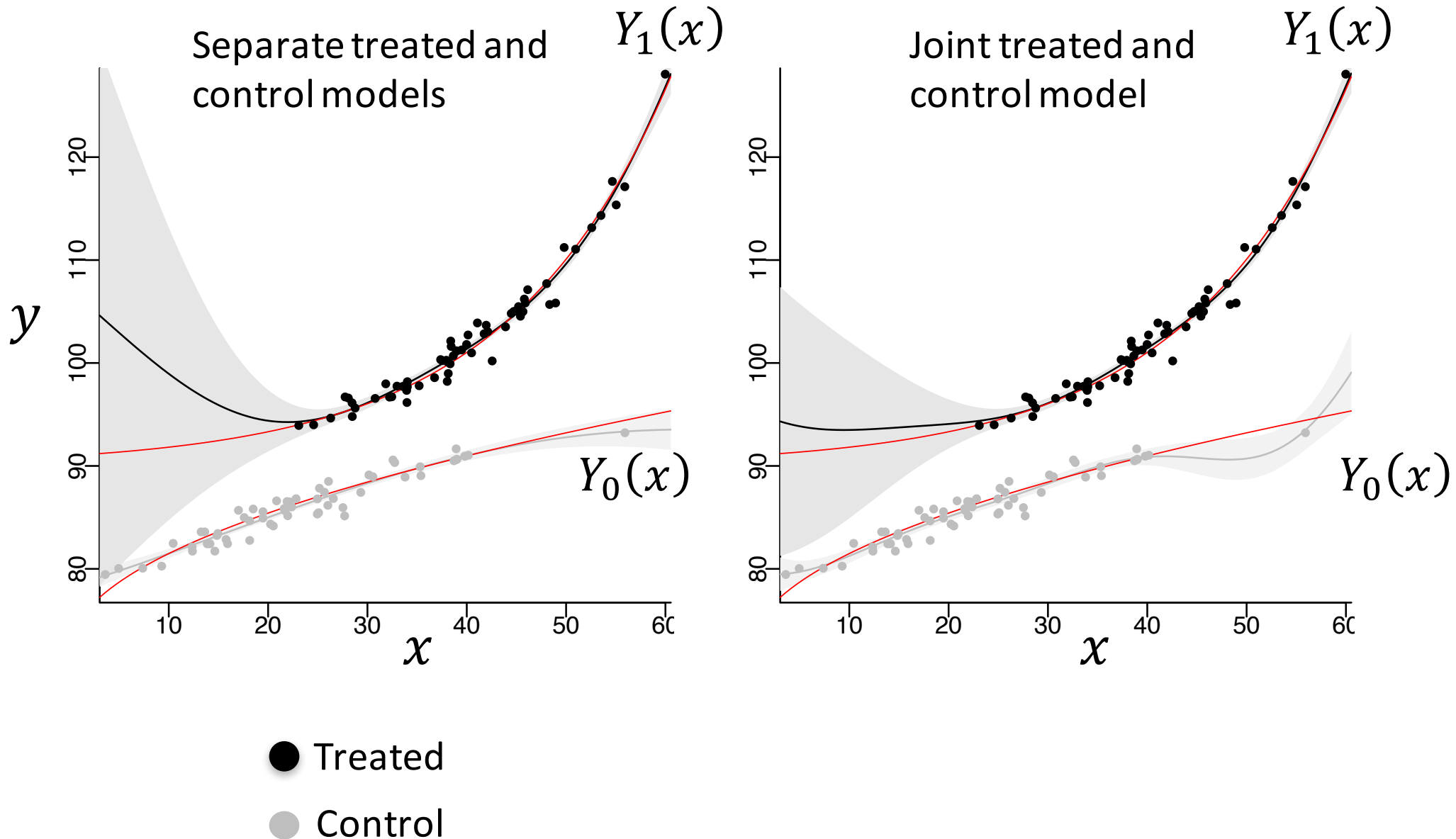
$$\hat{\gamma} = \gamma + \delta \frac{\mathbb{E}[xt]\mathbb{E}[x^2] - \mathbb{E}[t^2]\mathbb{E}[x^2t]}{\mathbb{E}[xt]^2 - \mathbb{E}[x^2]\mathbb{E}[t^2]}$$

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

Covariate adjustment with non-linear models

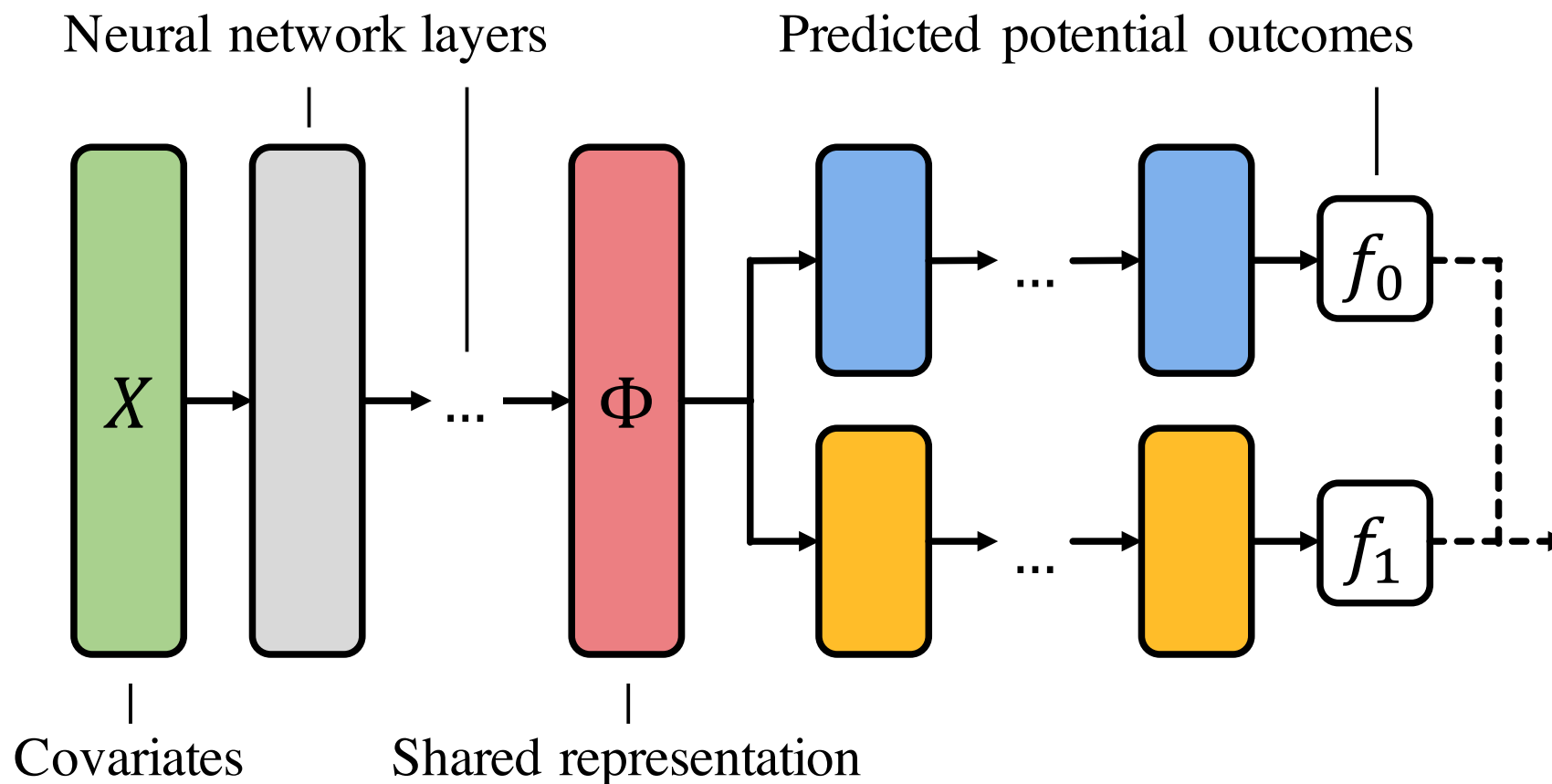
- Random forests and Bayesian trees
Hill (2011), Athey & Imbens (2015), Wager & Athey (2015)
- Gaussian processes
Hoyer et al. (2009), Zigler et al. (2012)
- Neural networks
Beck et al. (2000), Johansson et al. (2016), Shalit et al. (2016), Lopez-Paz et al. (2016)

Example: Gaussian processes



Figures: Vincent Dorie & Jennifer Hill

Example: Neural networks



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

Matching

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

Matching

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



Obama, had he gone to law school

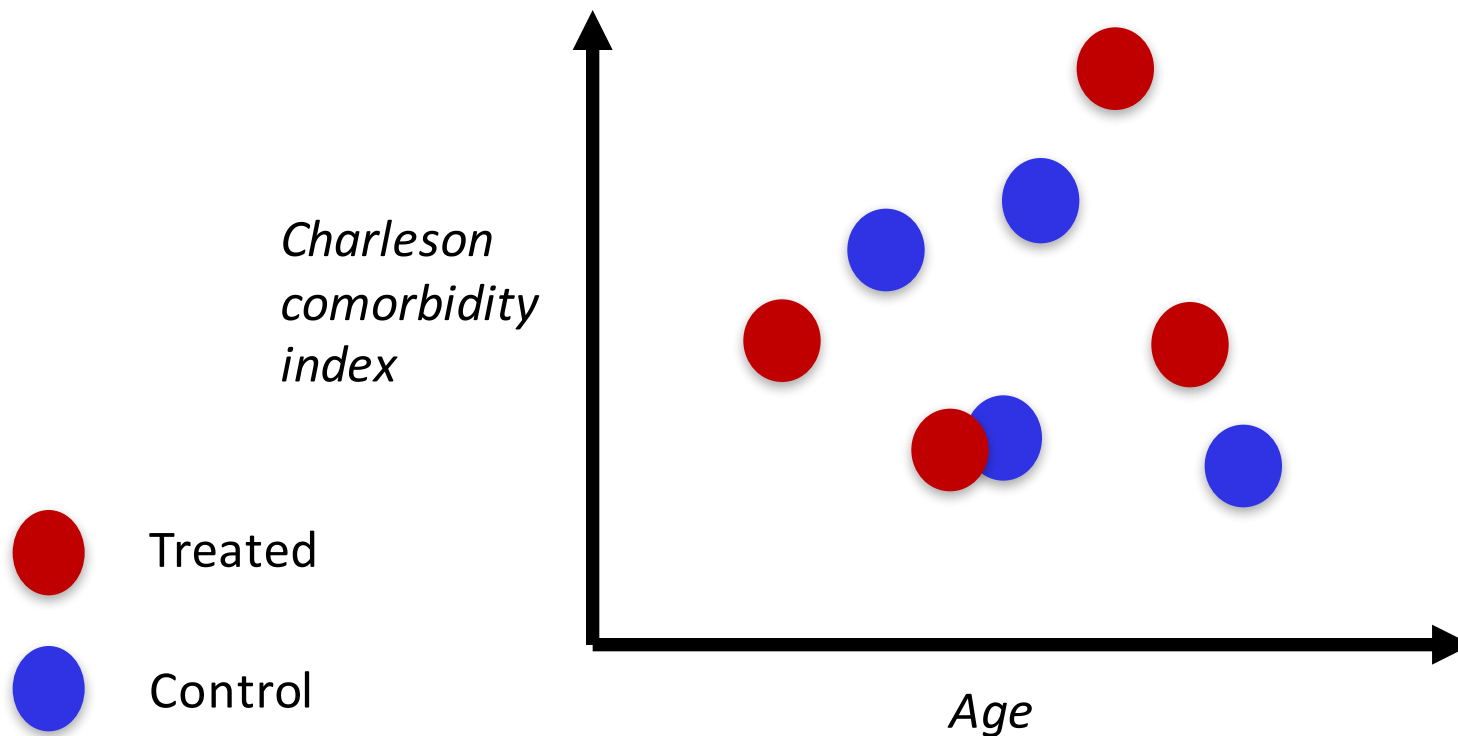


Obama, had he gone to business school

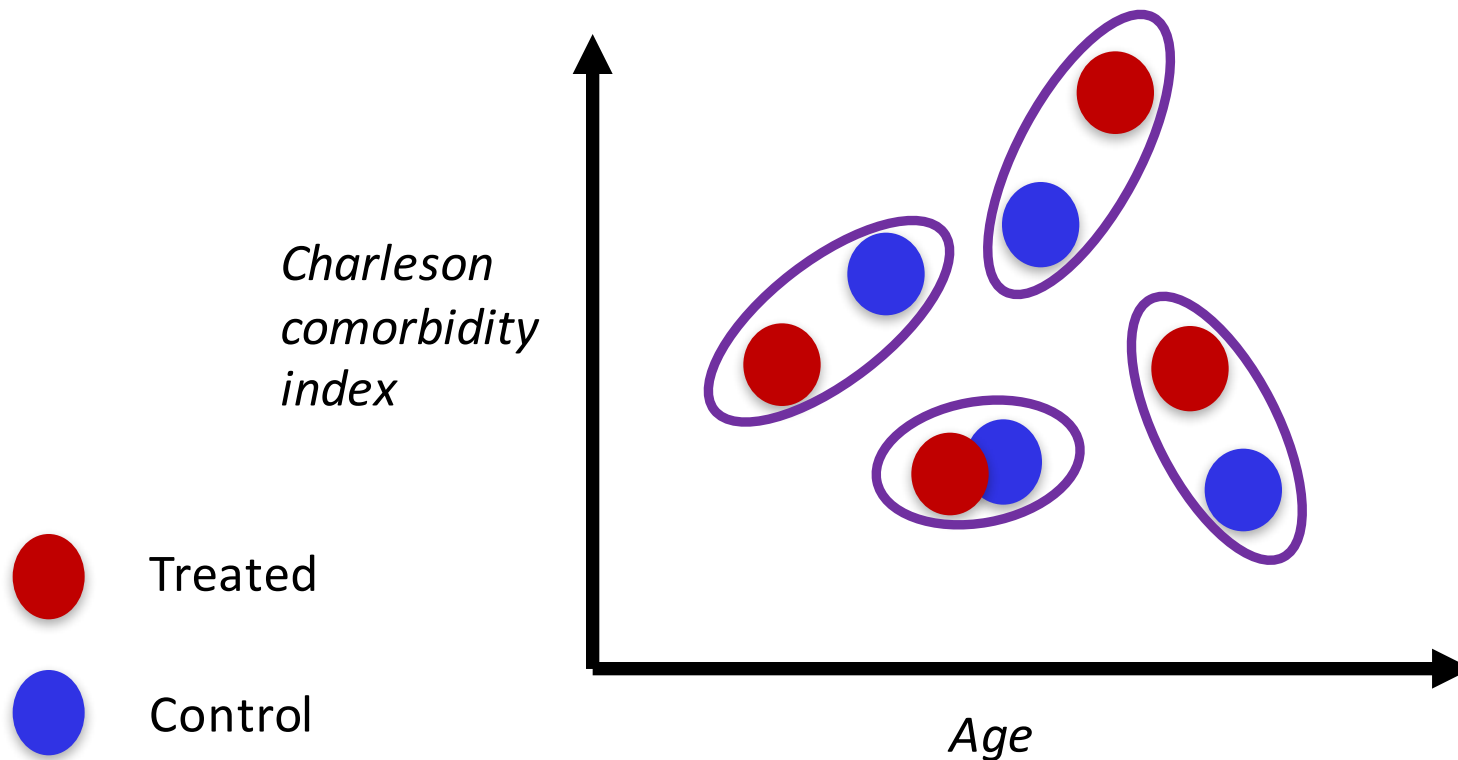
Matching

- Find each unit'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 $j(i) = \underset{j \text{ s.t. } t_j \neq t_i}{\operatorname{argmin}} d(x_j, x_i)$
 $j(i)$ 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

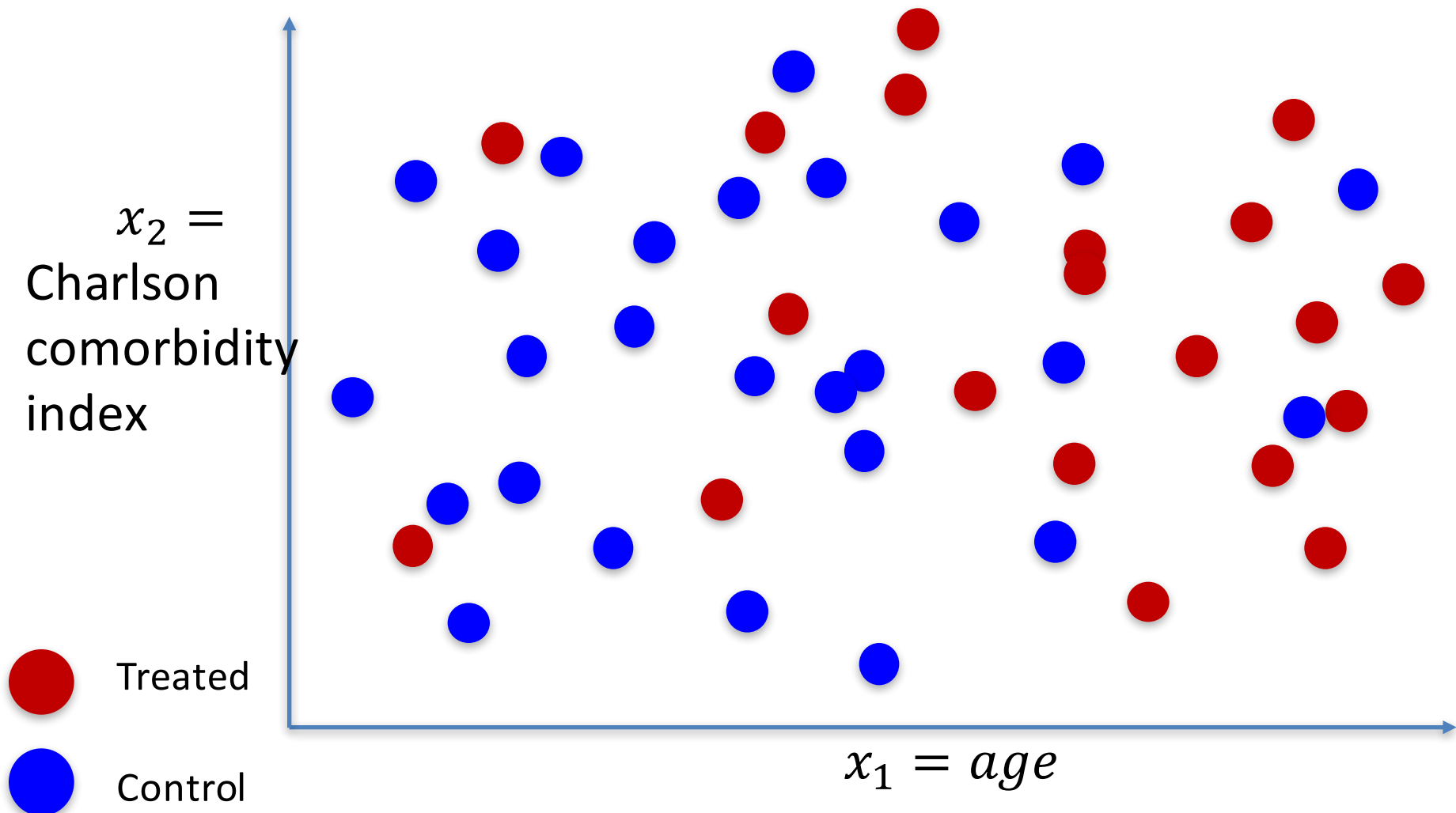
Propensity scores

- Tool for estimating ATE
- Basic idea: turn observational study into a pseudo-randomized trial by re-weighting samples, similar to importance sampling

Inverse propensity score re-weighting

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

reweighted control *reweighted treated*



Propensity score

- Propensity score: $p(T = 1|x)$,
using machine learning tools
- 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 $(x_1, t_1, y_1), \dots, (x_n, t_n, y_n)$

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 $(x_1, t_1, y_1), \dots, (x_n, t_n, y_n)$

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 $(x_1, t_1, y_1), \dots, (x_n, t_n, y_n)$

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 $(x_1, t_1, y_1), \dots, (x_n, t_n, y_n)$

1. Randomized trial $p = 0.5$

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

Propensity scores – algorithm

Inverse probability of treatment weighted estimator

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

1. Randomized trial $p = 0.5$

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

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

Propensity scores - derivation

- Recall average treatment effect:

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

- We only have samples for:

$$\mathbb{E}_{x \sim p(x|T=1)} [\mathbb{E} [Y_1 | x, T = 1]]$$

$$\mathbb{E}_{x \sim p(x|T=0)} [\mathbb{E} [Y_0 | x, T = 0]]$$

Propensity scores - derivation

- We only have samples for:

$$\mathbb{E}_{x \sim p(x|T=1)} [\mathbb{E}[Y_1|x, T=1]]$$

$$\mathbb{E}_{x \sim p(x|T=0)} [\mathbb{E}[Y_0|x, T=0]]$$

Propensity scores - derivation

- We only have samples for:

$$\mathbb{E}_{x \sim p(x|T=1)} [\mathbb{E}[Y_1|x, T=1]]$$

$$\mathbb{E}_{x \sim p(x|T=0)} [\mathbb{E}[Y_0|x, T=0]]$$

- We need to turn $p(x|T=1)$ into $p(x)$:

$$p(x|T=1) \cdot \quad ? \quad = p(x)$$

Propensity scores - derivation

- We only have samples for:

$$\mathbb{E}_{x \sim p(x|T=1)} [\mathbb{E}[Y_1|x, T=1]]$$

$$\mathbb{E}_{x \sim p(x|T=0)} [\mathbb{E}[Y_0|x, T=0]]$$

- We need to turn $p(x|T=1)$ into $p(x)$:

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

Propensity score

Propensity scores - derivation

- We only have samples for:

$$\mathbb{E}_{x \sim p(x|T=1)} [\mathbb{E}[Y_1|x, T=1]]$$

$$\mathbb{E}_{x \sim p(x|T=0)} [\mathbb{E}[Y_0|x, T=0]]$$

- We need to turn $p(x|T=0)$ into $p(x)$:

$$p(x|T=0) \cdot \frac{p(T=0)}{p(T=0|x)} = p(x)$$

Propensity score

- 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:

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

- We can approximate this empirically as:

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

(similarly for $t_i=0$)

Problems with 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

Many more ideas and methods

- Natural experiments & regression discontinuity
- Instrumental variables

Many more ideas and methods – Natural experiments

- Does stress during pregnancy affect later child development?
- Confounding: genetic, mother personality, economic factors...
- Natural experiment: the Cuban missile crisis of October 1962. Many people were afraid a nuclear war is about to break out.
- Compare children who were in utero during the crisis with children from immediately before and after

Many more ideas and methods – Instrumental variables

- Informally: a variable which affects treatment assignment but not the outcome
- Example: are private schools better than public schools?
- Confounding: different student population, different teacher population
- Can't force people which school to go to

Many more ideas and methods – Instrumental variables

- Informally: a variable which affects treatment assignment but not the outcome
- Example: are private schools better than public schools?
- Can't force people which school to go to
- *Can randomly give out vouchers to some children, giving them an opportunity to attend private schools*
- *The voucher assignment is the instrumental variable*

Summary

- Two approaches to use machine learning for causal inference:
 1. Predict outcome given features and treatment, then use resulting model to impute counterfactuals (*covariate adjustment*)
 2. Predict treatment using features (*propensity score*), then use to reweight outcome or stratify the data
- Causal graphs important for thinking through whether problem is setup appropriately and whether assumptions hold