Machine Learning for Healthcare

HST.956, 6.S897

Lecture 15: Causal Inference Part 2

David Sontag

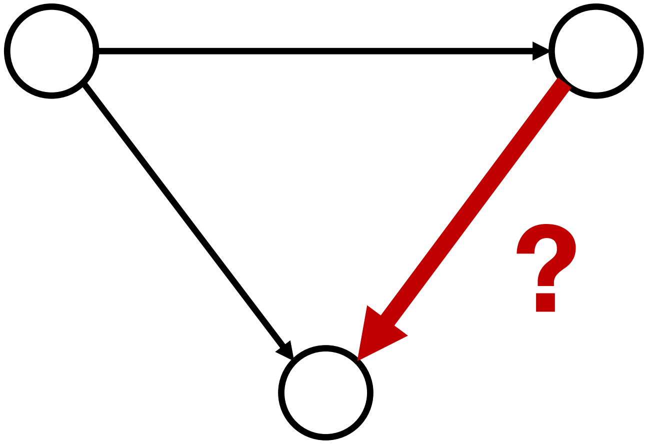
CSAIL

INSTITUTE FOR MEDICAL
ENGINEERING & SCIENCE

HEALTH SCIENCES
& TECHNOLOGY

Acknowledgement: adapted from slides by Uri Shalit (Technion)

Reminder: Causal inference



Patient, X

(including all confounding factors)

Intervention, T

(e.g. medication, procedure)

Outcome, Y

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

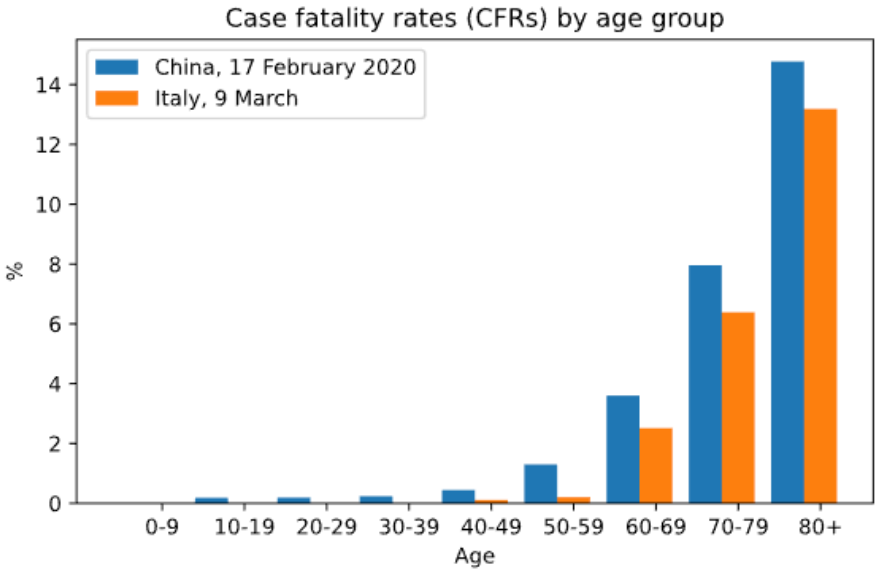
Causal inference for COVID19

-

Causal inference for COVID19

- Example (simplified; for educational purposes only)
  - Understanding case fatality rates (CFR)

    Paradox: CFR in Italy reported at 4.3% and CFR in China reported at 2.3%. Yet:



Courtesy of Julius von Kuegelgen & Luigi Gresele. Used with permission.

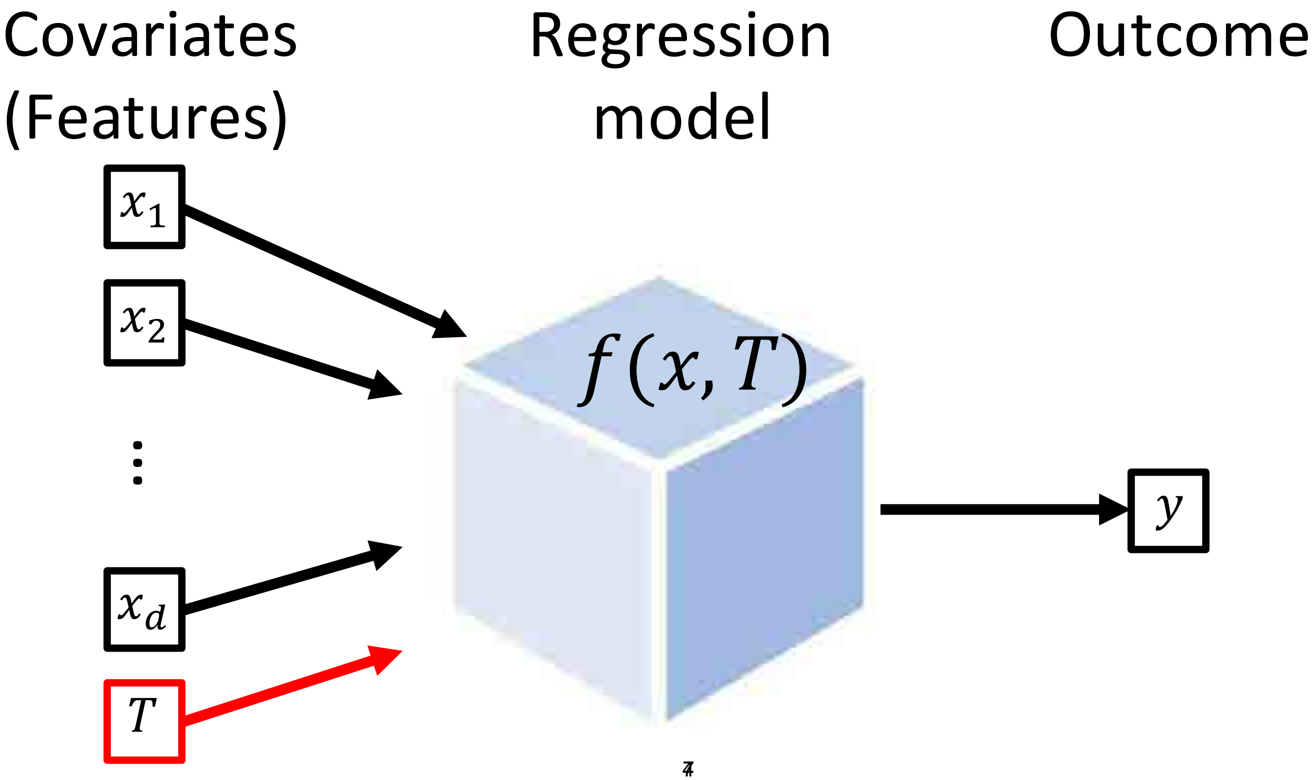
Two common approaches for counterfactual inference

Covariate adjustment

Propensity scores

Covariate adjustment (reminder)

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



Covariates (Features)

Regression model

Outcome

Covariate adjustment (reminder)

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

- 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:

$${}_1(x) - Y_0(x)] =$$

Covariate ad ust e t with ear odes

- Assume that:

B d ressure **age** **medicati**

$$_t(\quad) = \qquad + \gamma \cdot \quad + \epsilon_t$$

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

- The :

$$\overline{\quad}$$

$$CATE(x) := \mathbb{E}[\,_1(\quad) \qquad _0(x)] =$$

$$\mathbb{E}[\cancel{(}\qquad + \gamma + \epsilon_1) - \cancel{(}\qquad + \epsilon_0)] =$$

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

Covariate ad ustment wt linear modes

- Assume t at:

B ood pressure age med cat o

$$_t(\quad) = \qquad + \gamma \cdot \quad + \epsilon_t$$

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

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

- For causa nference, need to est mate well, not $_t(\quad)$ - **Identf icat on, not predict on**
- *Ma or di ere ce betwee ML and statistics*

W at appens true mode s not linear?

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

$$Y_t(\;) = \quad + \quad \cdot t + \quad \cdot x^2$$

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

- Hy≠pot esized model:

$$\widehat{Y}_t(\;) \quad \hat{\ } x + \hat{\ } \cdot$$

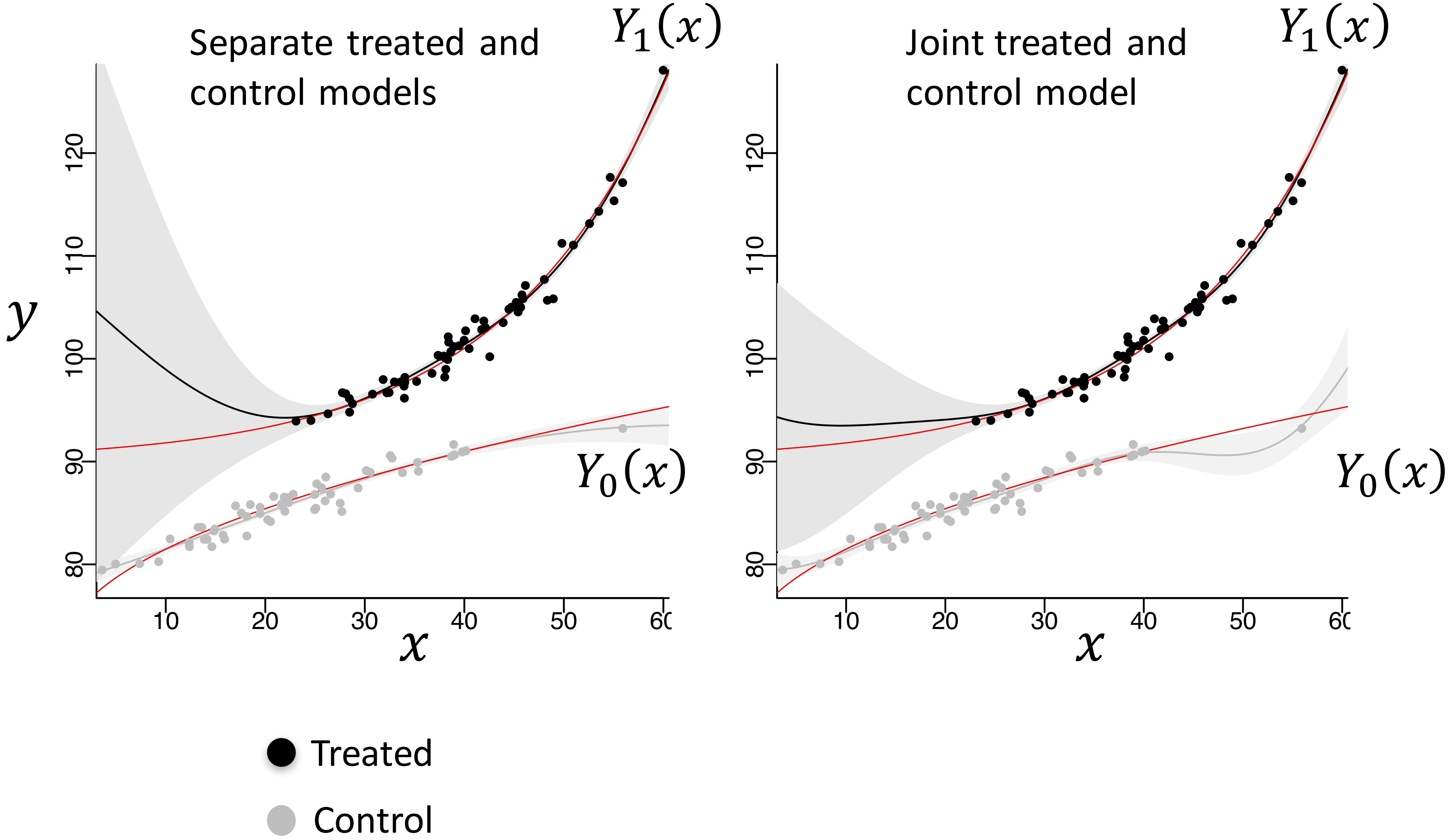
$$\hat{\ } = \quad + \frac{\mathbb{E}[\quad]\mathbb{E}[\ ^2] - \mathbb{E}[\ ^2]\mathbb{E}[\ ^2]}{\mathbb{E}[\quad]^2 - \mathbb{E}[\ ^2]\mathbb{E}[t^2]}$$

Dependig δ, ca be made t be arb trar y arge r
s al!

Covariate adjustment with non-linear models

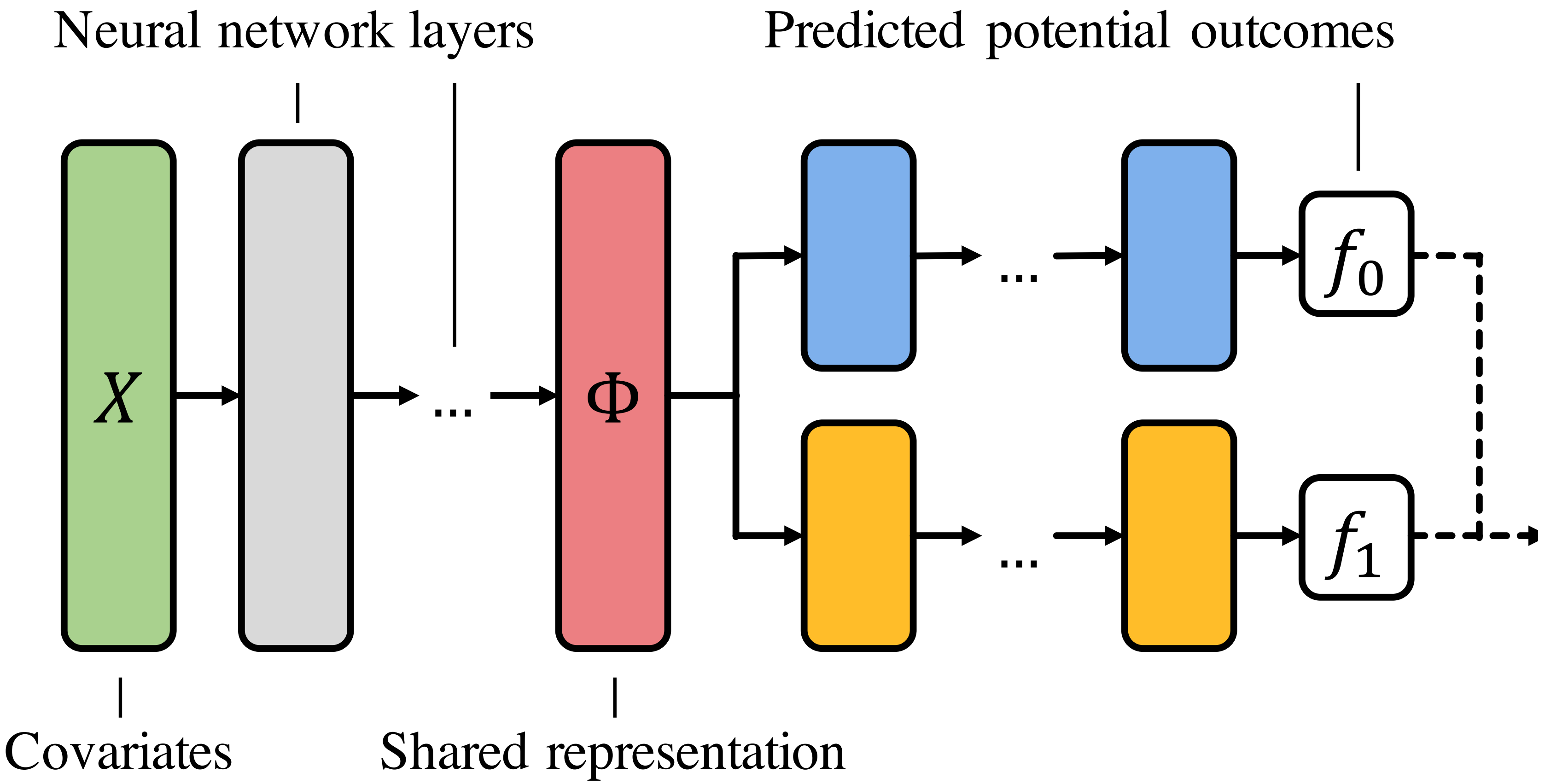
- Random forests and Bayesian trees
  Hill (2011), Athey & Imbens (2015), Wager & Athey (2015)
- Gaussian processes
  Hoyer et al. (2009), Zigler et al. (2012)
- Neural networks
  Beck et al. (2000), Johansson et al. (2016), Shalit et al. (2016), Lopez-Paz et al. (2016)

Example: Gaussian processes


Separate treated and control models

Joint treated and control model

Example: Neural networks

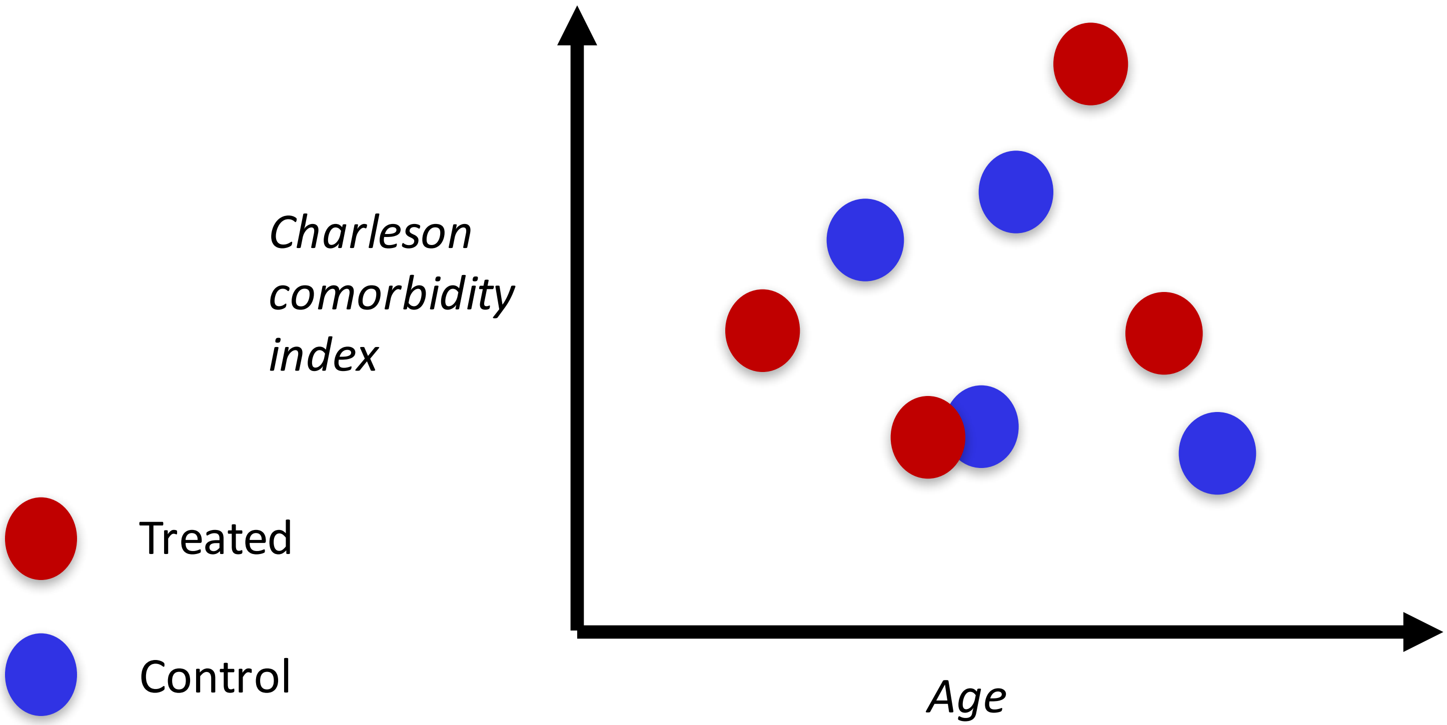


Shalit, Johansson, Sontag. *Estimating Individual Treatment Effect: Generalization Bounds and Algorithms*. ICML, 2017

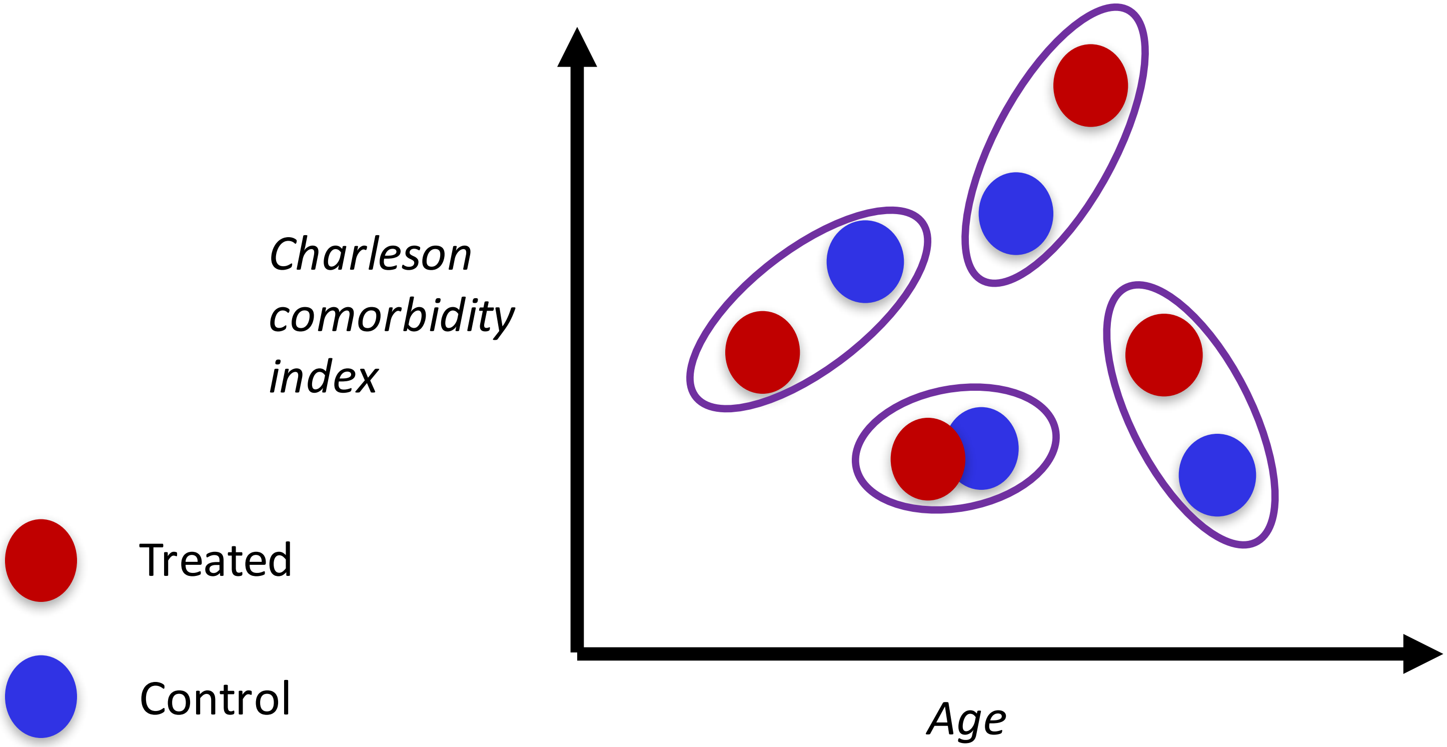
Matching

- Find each unit’s long-lost counterfactual identical twin, check up on his outcome
- 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 $j(i) = \underset{j\ s.t.\ t_j \neq t_i}{\operatorname{argmin}}\ d(x_j, x_i)$

  $j(i)$ 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:

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

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

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

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 $y_{NN_t(\ \ )}$ 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

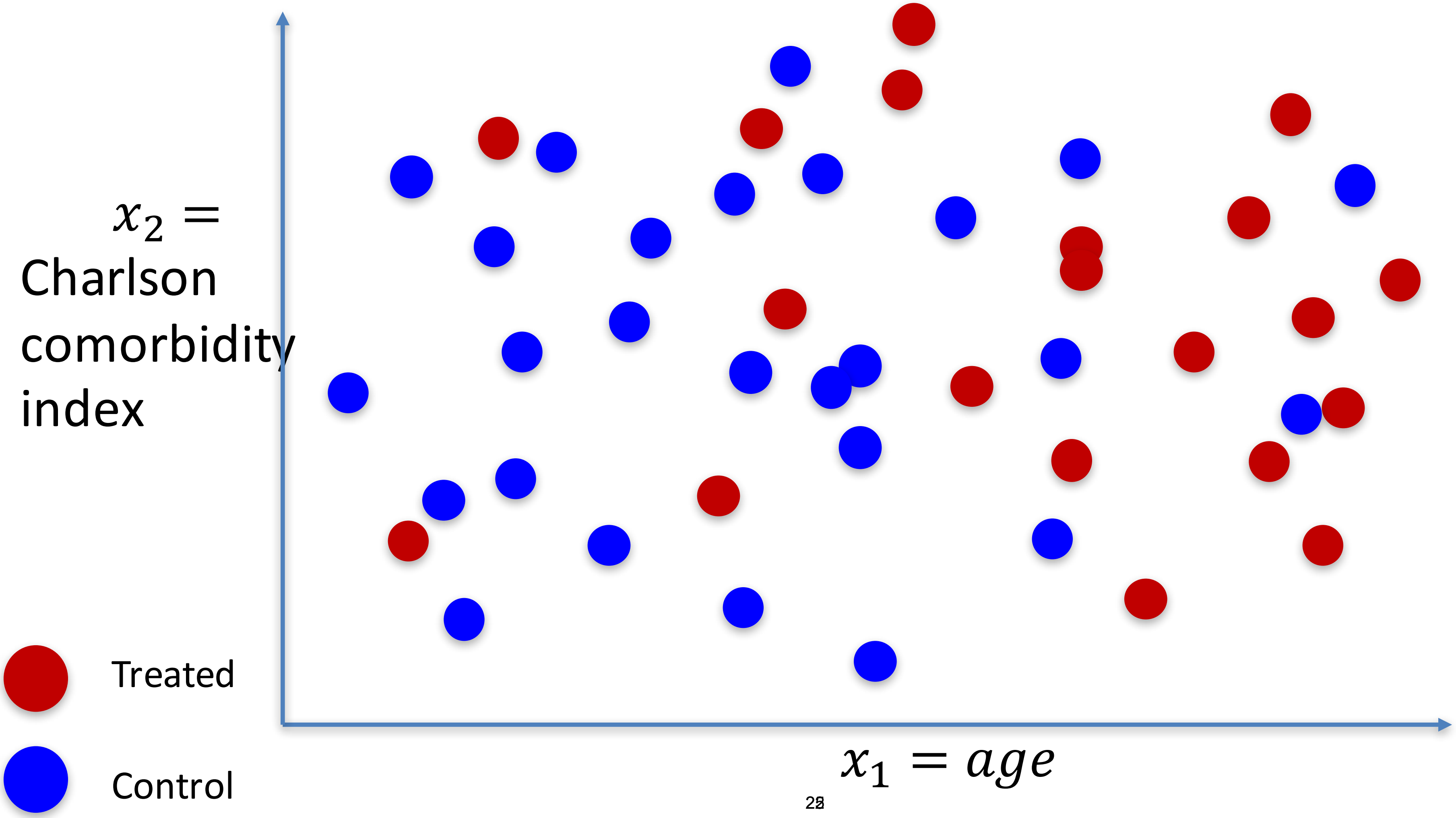
Propensity scores

Propensity scores

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

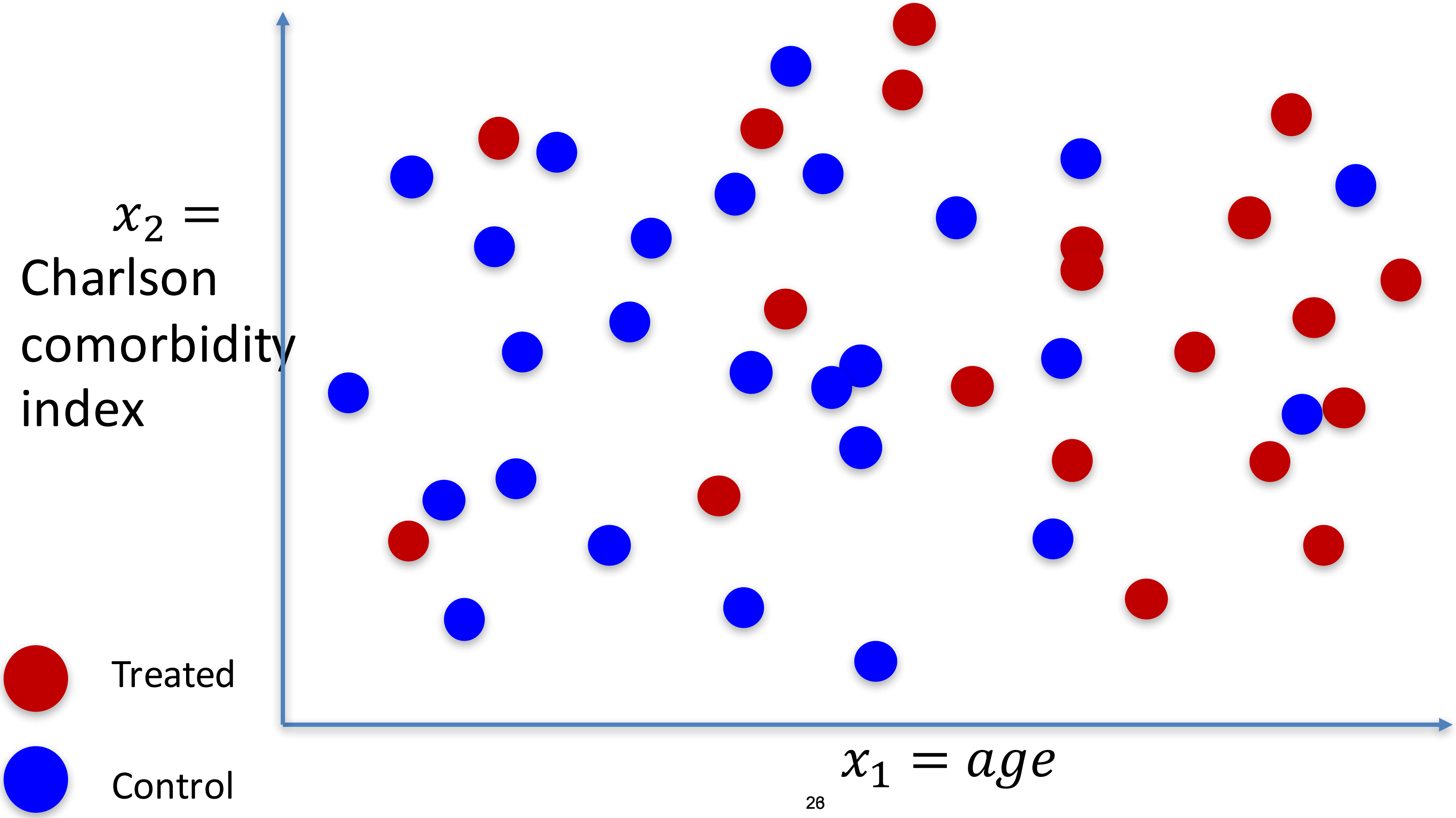
Inverse propensity score re-weighting

$$\underset{control}{p(x|t=0)} \neq \underset{treated}{p(x|t=1)}$$



Inverse propensity score re-weighting

$$\underbrace{p(x|t=0)\cdot w_0(x)}_{reweighted\ control} \approx \underbrace{p(x|t=1)\cdot w_1(x)}_{reweighted\ treated}$$



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

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 $p = 0.5$

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

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

1. Randomized trial $p = 0.5$

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

Propensity scores - 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:

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

- We can approximate this empirically as:

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

(similarly for t_i=0)

Problems with 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
  - 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/1810.02894