6.S979 Topics in Deployable ML, Fall 2019

Causal Inference and Predicting Counterfactuals

David Sontag

Acknowledgement: many slides made by Uri Shalit for our ICML 2016 tutorial

CSAIL

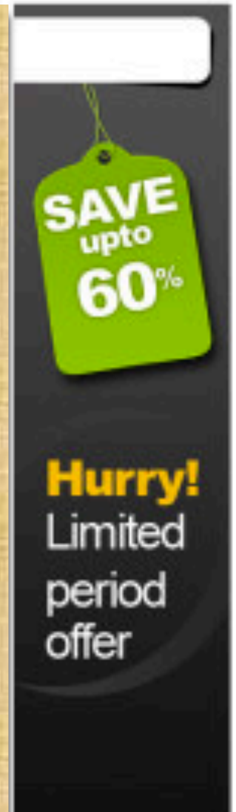
INSTITUTE FOR MEDICAL ENGINEERING & SCIENCE


Is smoking
dangerous?

Do stricter gun laws lead to safer communities?

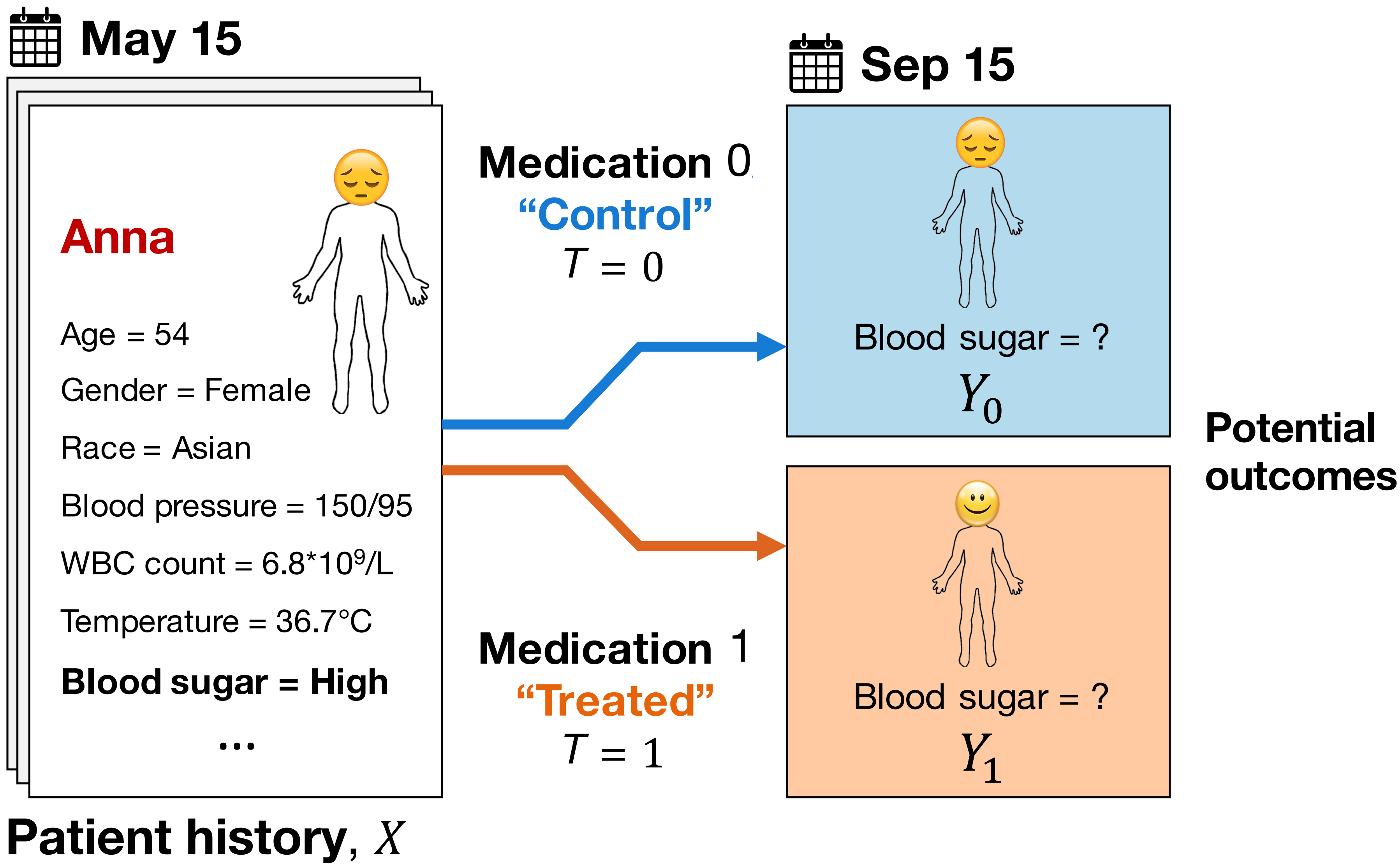
Is pre-kindergarten beneficial for children?

Will running an ad-campaign increase sales?



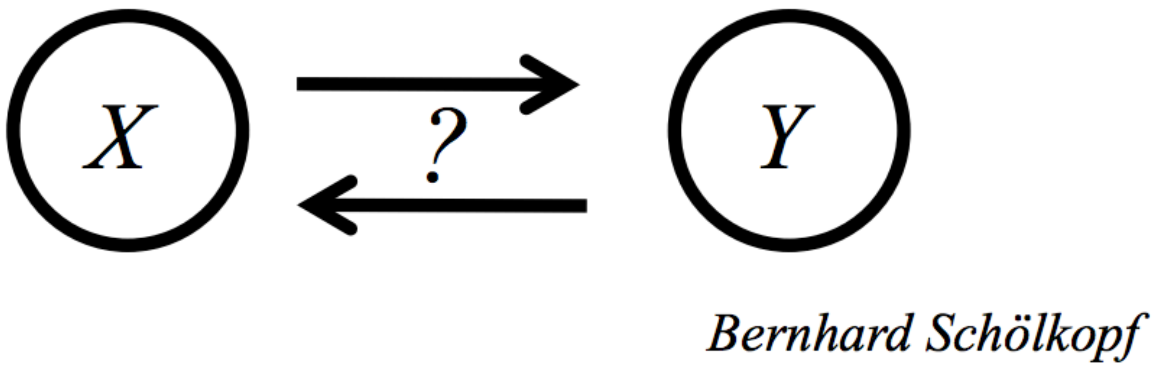
Did a company discriminate against job applicants?

Which medication to prescribe?



Machine learning "causals" that we won't discuss:

- Identifying causal direction between two variables:



- Assumptions on noise process
- Work by Schölkopf, Janzing, Guyon, Mooij, Peters, Geiger, Lopez-Paz and others

Machine learning "causals" that we won't discuss:

- Learning causal graph structure from data:
  - Distinguishes between direct and indirect effects
  - Makes different set of assumptions, such as "faithfulness"
- Bühlmann, Geng, Maathuis, Pearl, Meinshausen, Tsamardinos...

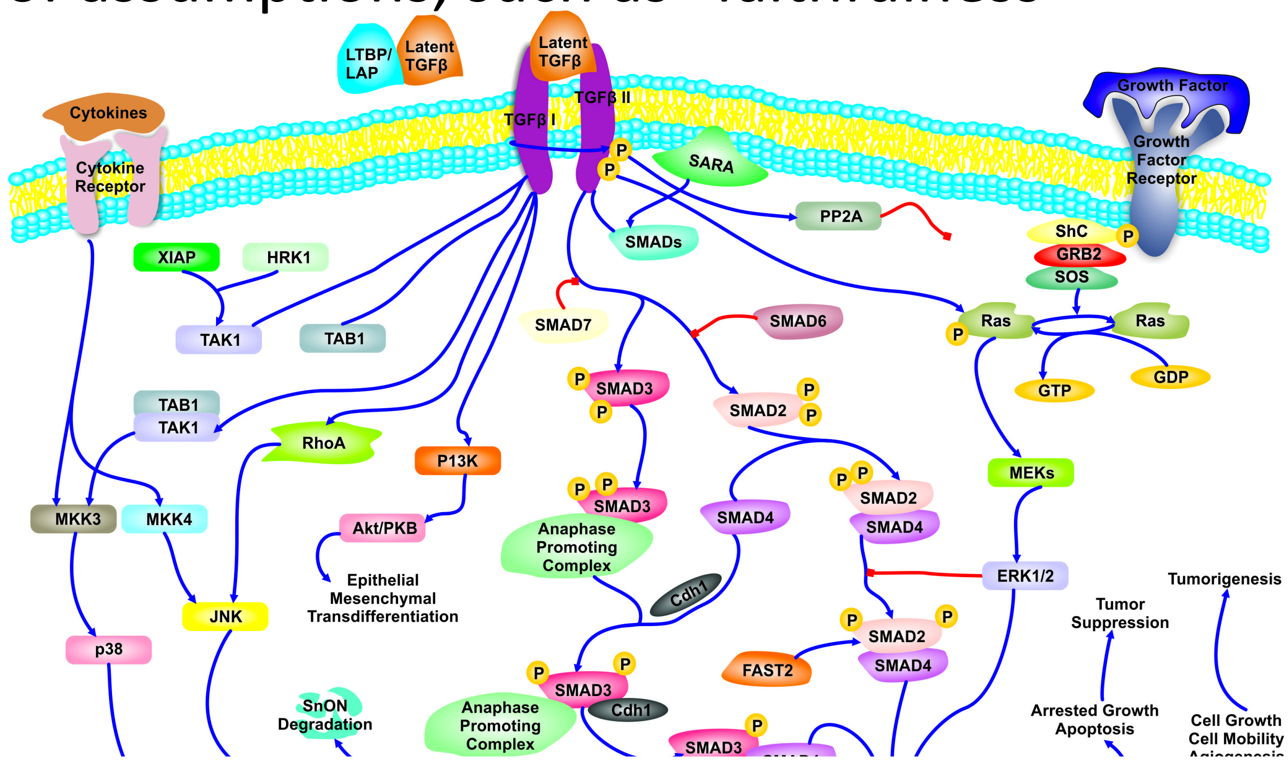
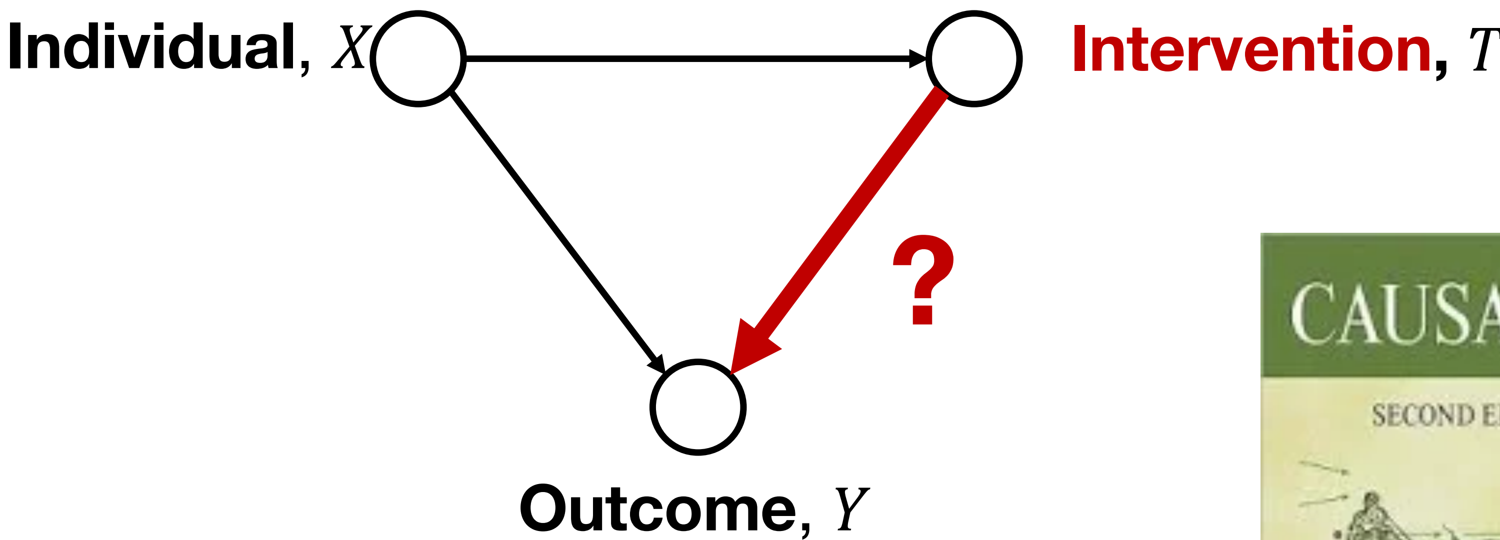
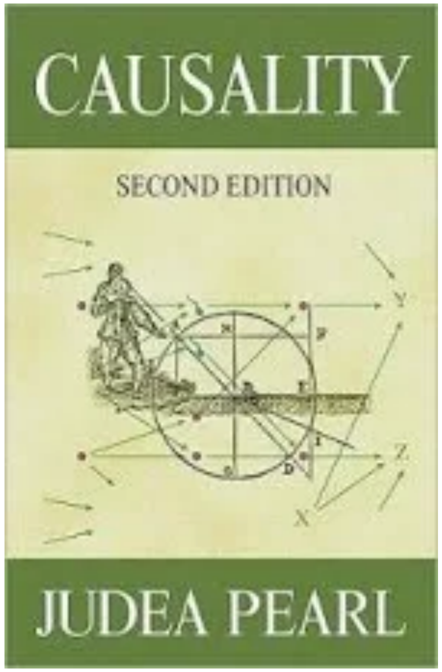


Image from: http://www.mensxmachina.org/causalpath/state.html

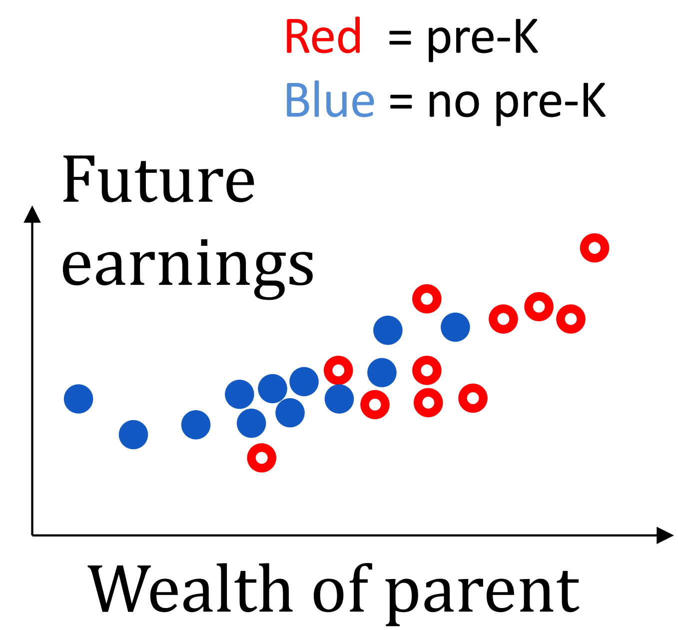
Formalization using language of causality





Key challenge: bias in data

- Here, wealth of parent is a **confounder**. If not corrected for, would obtain a biased estimate of causal effect
- If *no* young patients treated, lack **treatment group overlap**—estimation impossible without strong assumptions



Potential Outcomes Framework (Rubin-Neyman Causal Model)

- 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:

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

Potential Outcomes Framework (Rubin-Neyman Causal Model)

- 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***"

- Observed factual outcome:

$$y_i = t_i Y_1(x_i) + (1 - t_i) Y_0(x_i)$$

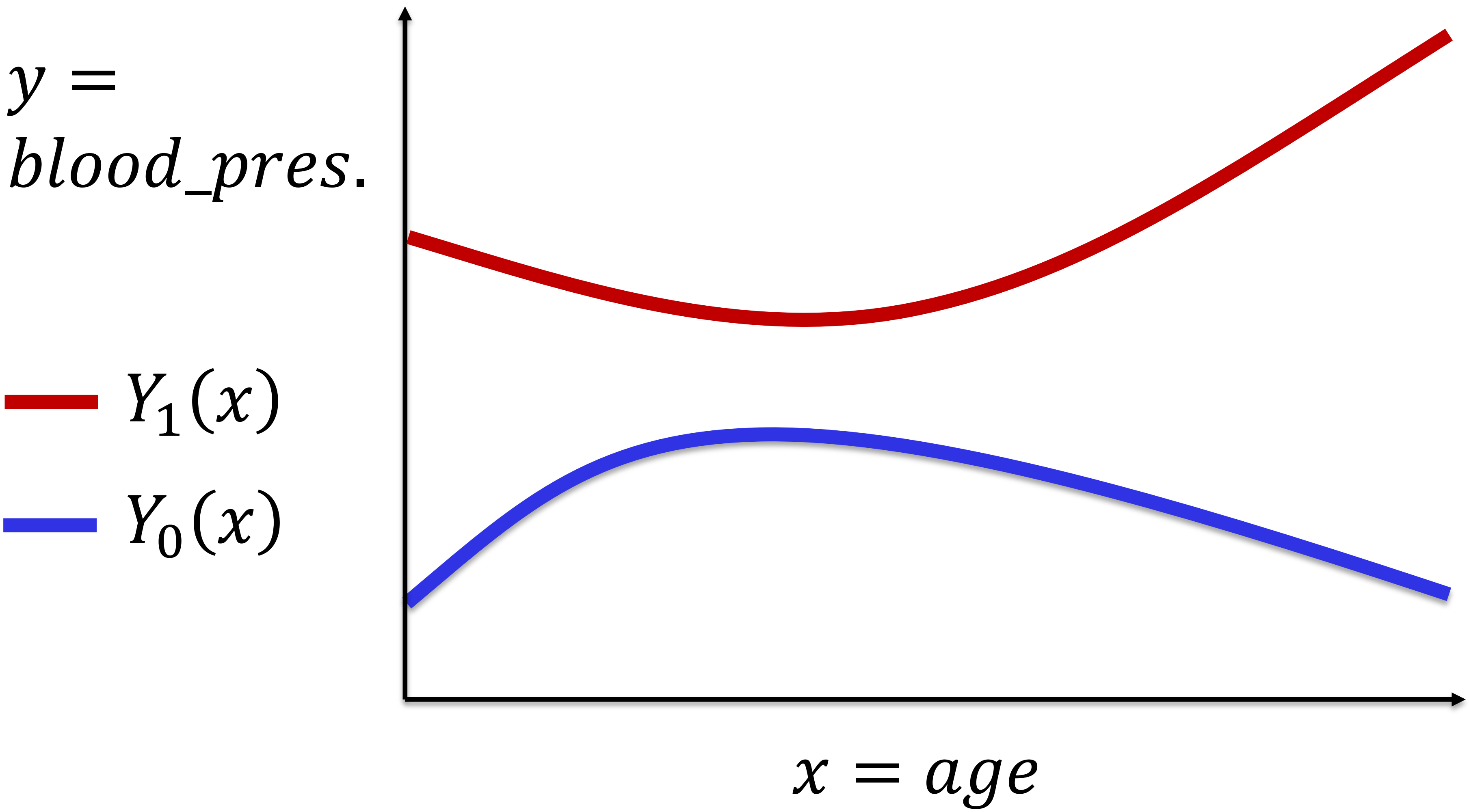
- Unobserved counterfactual outcome:

$$y_i^{CF} = (1 - t_i) Y_1(x_i) + t_i Y_0(x_i)$$

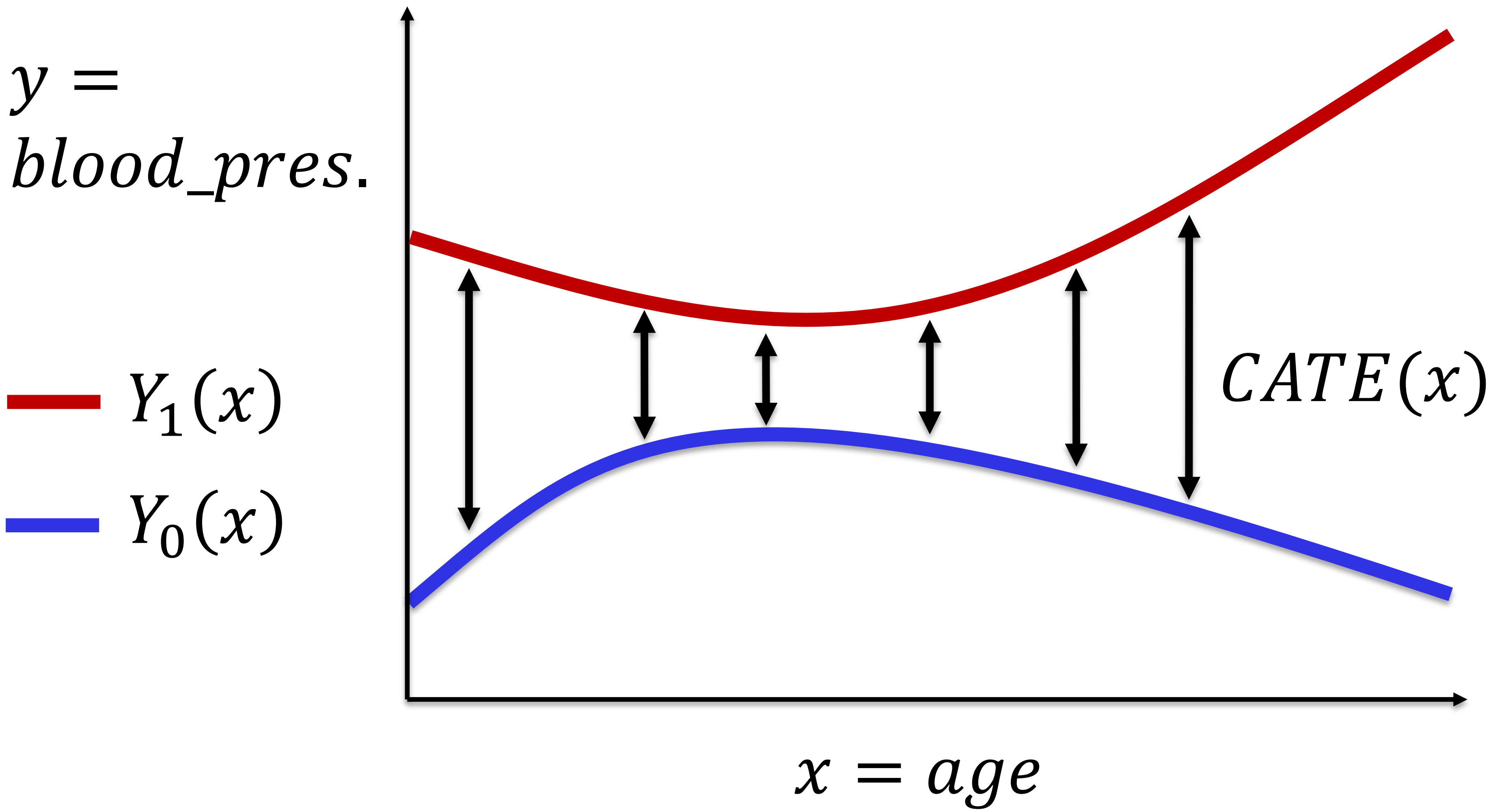
“The fundamental problem of causal inference”

We only ever observe one of the two outcomes

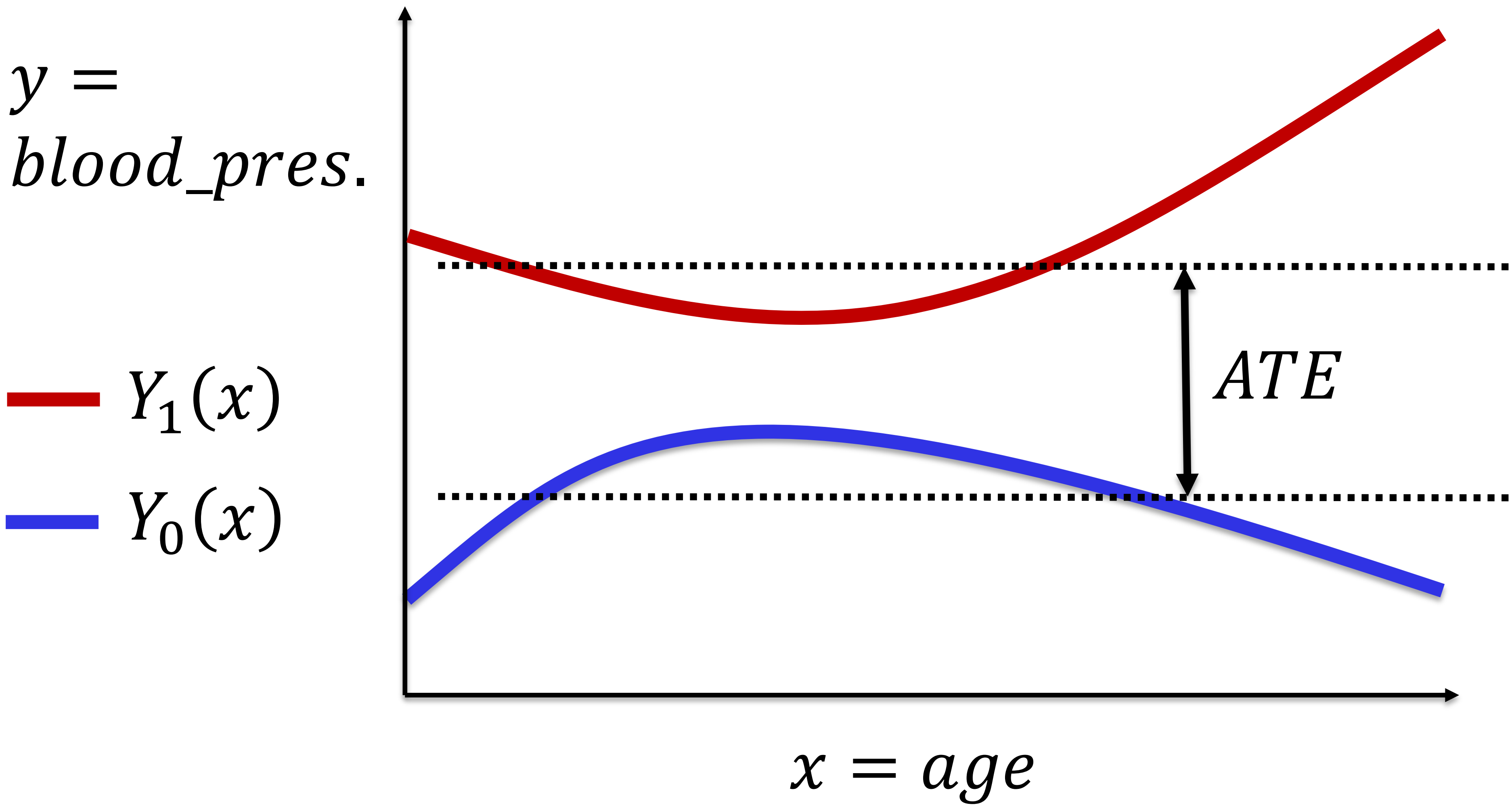
Example – Blood pressure and age



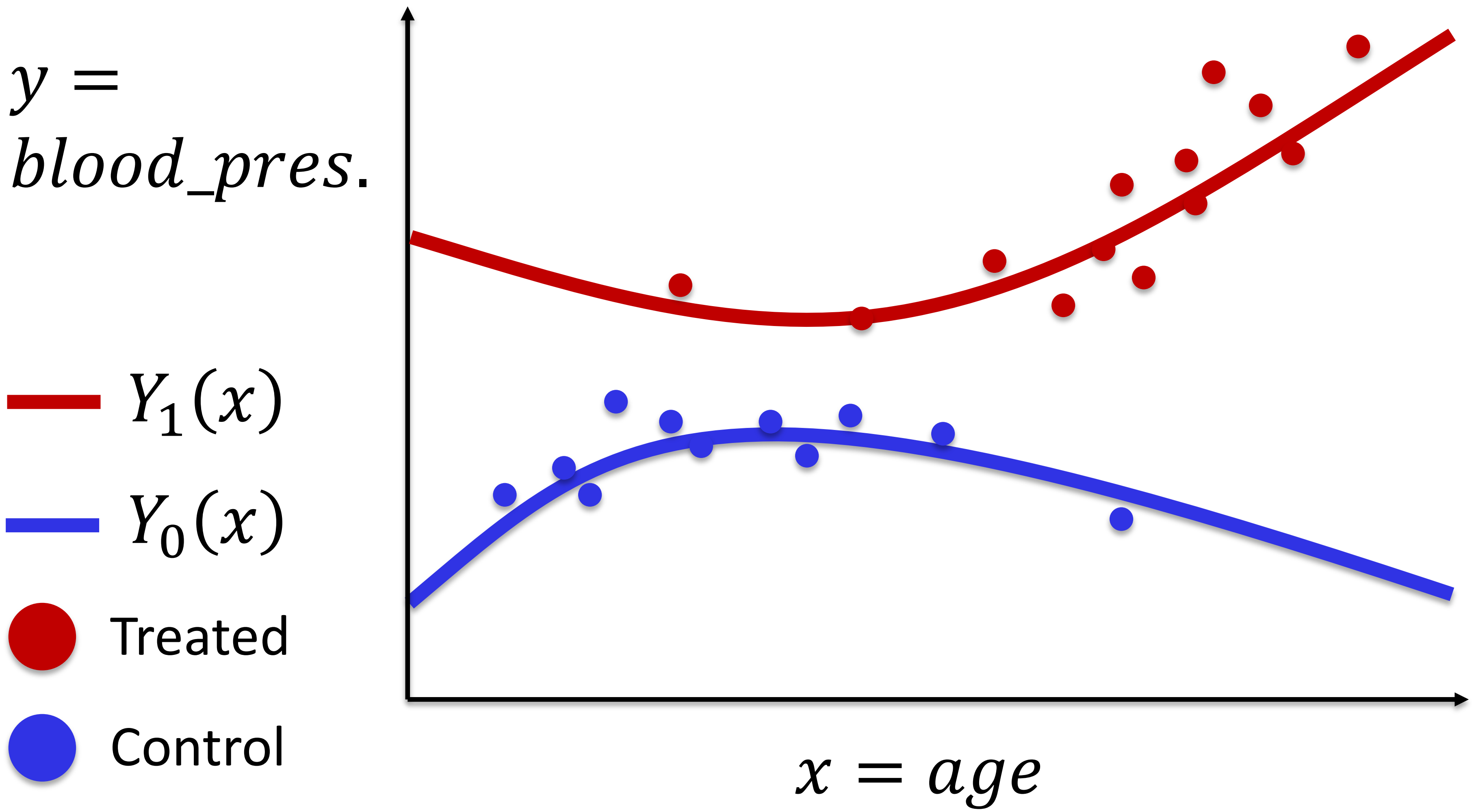
Blood pressure and age



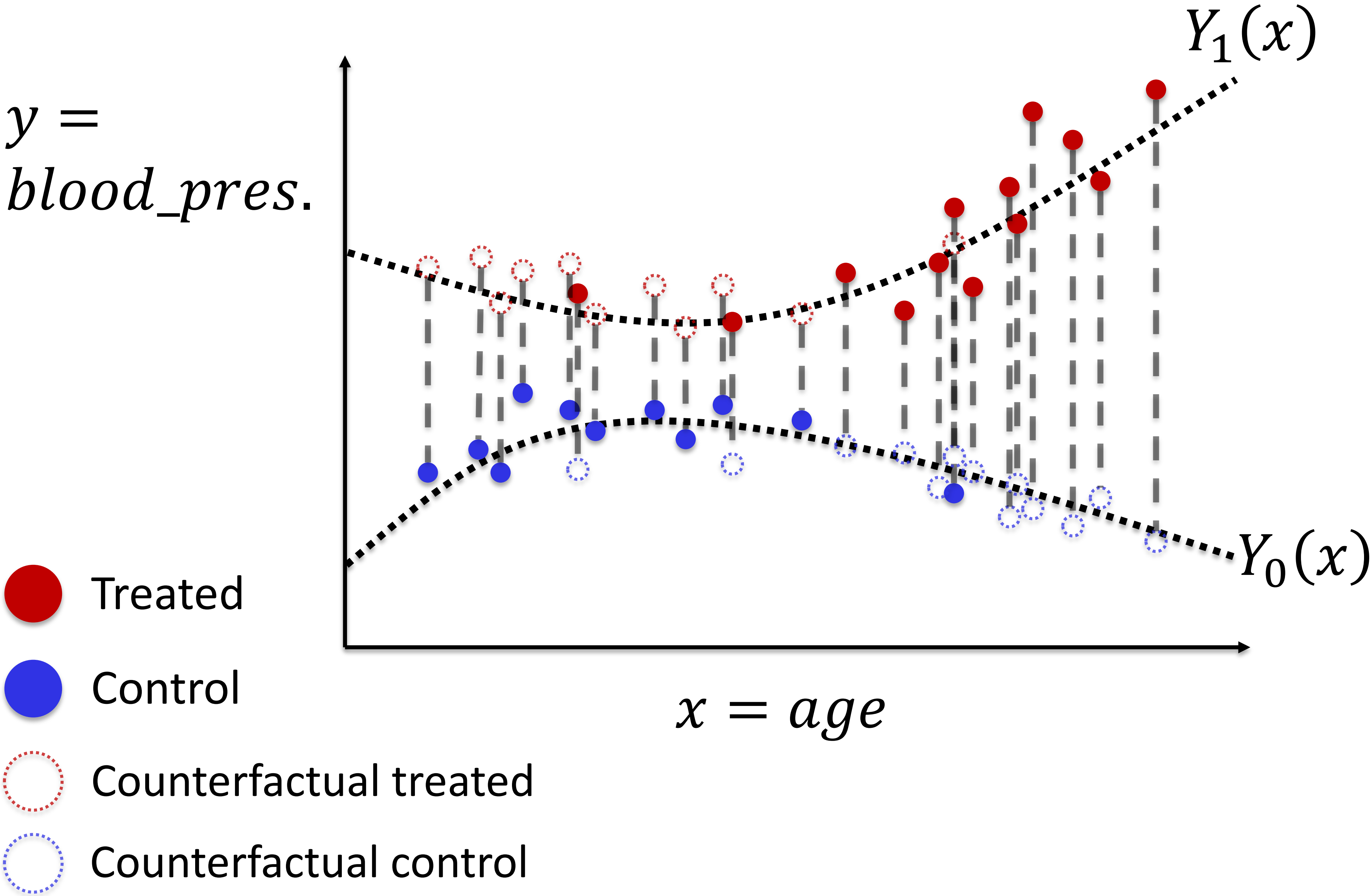
Blood pressure and age



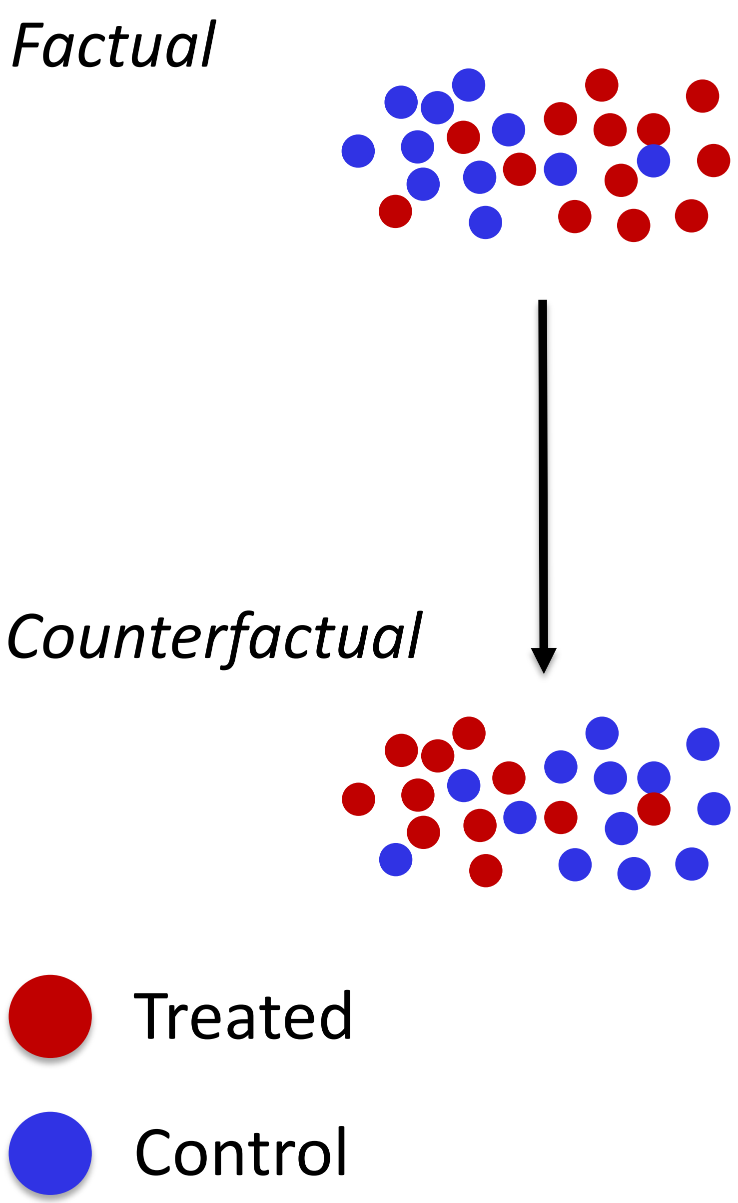
Blood pressure and age



Blood pressure and age



Connection to domain adaptation



$$p_F(x,t) = p_F(x)p_F(t|x)$$

the joint *factual* distribution over covariates and treatment assignment

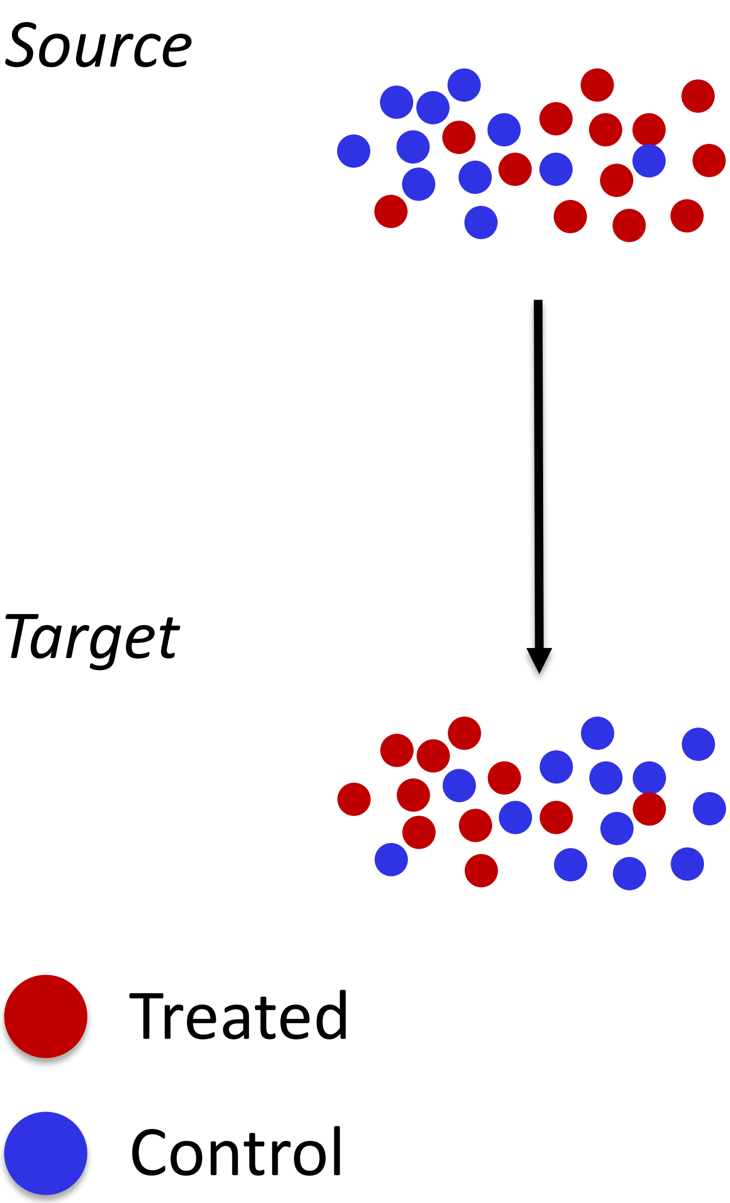
labeled y_i

$$p_{CF}(x,t) := p_F(x)p_F(1-t|x)$$

the joint *counterfactual* distribution over covariates and treatment assignment

unlabeled

Connection to domain adaptation



$p_{source}(x)$

the *source*

distribution over covariates

labeled

$p_{target}(x)$

the *target*

distribution over covariates

unlabeled

Typical assumption #1 – common support (overlap)

Y_0, Y_1: potential outcomes for control and treated

x: unit covariates (features)

T: treatment assignment

We assume:

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

Why is this a necessary assumption?

Note: in a randomized control trial (RCT) with two arms, p(T | X) = p(T) = 1/2

Typical assumption #2 – no unmeasured confounders

Y_0, Y_1: potential outcomes for control and treated

x: unit covariates (features)

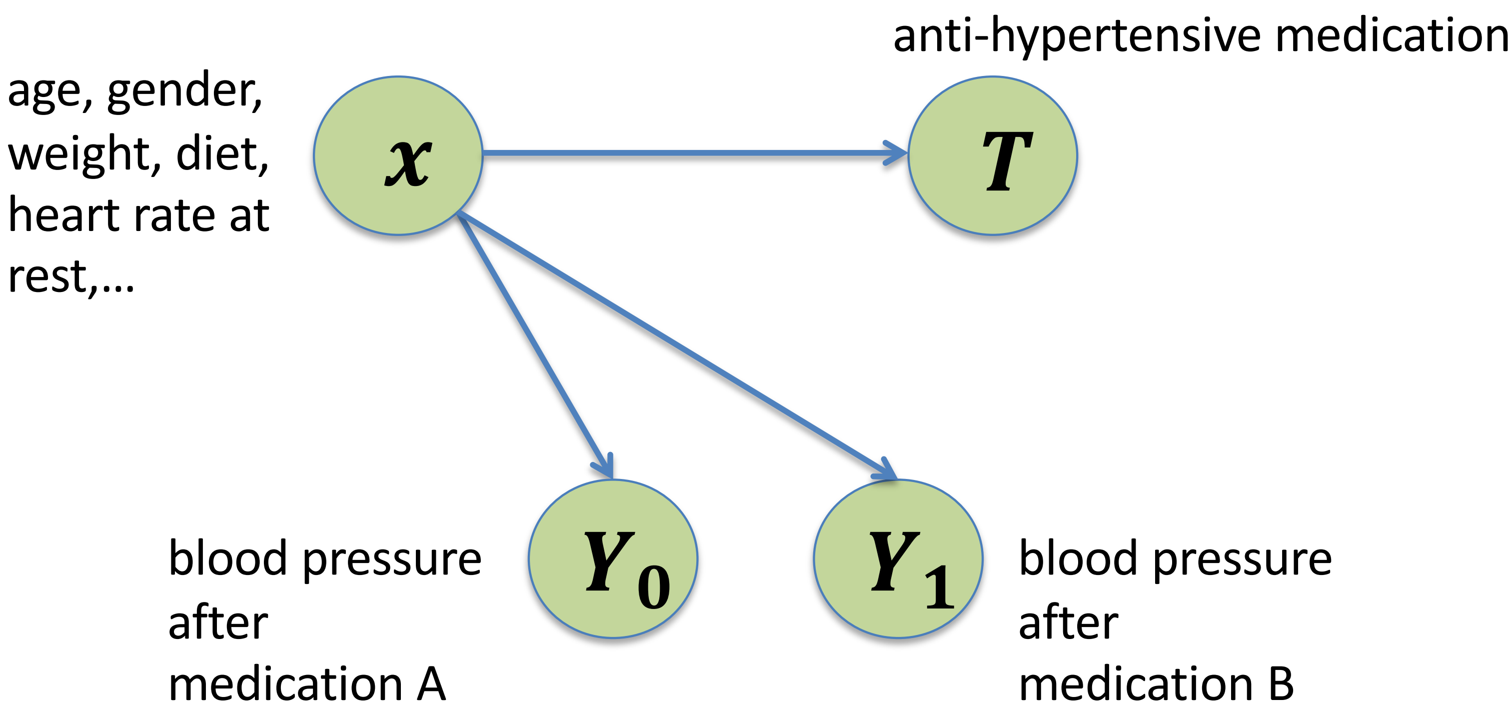
T: treatment assignment

We assume:

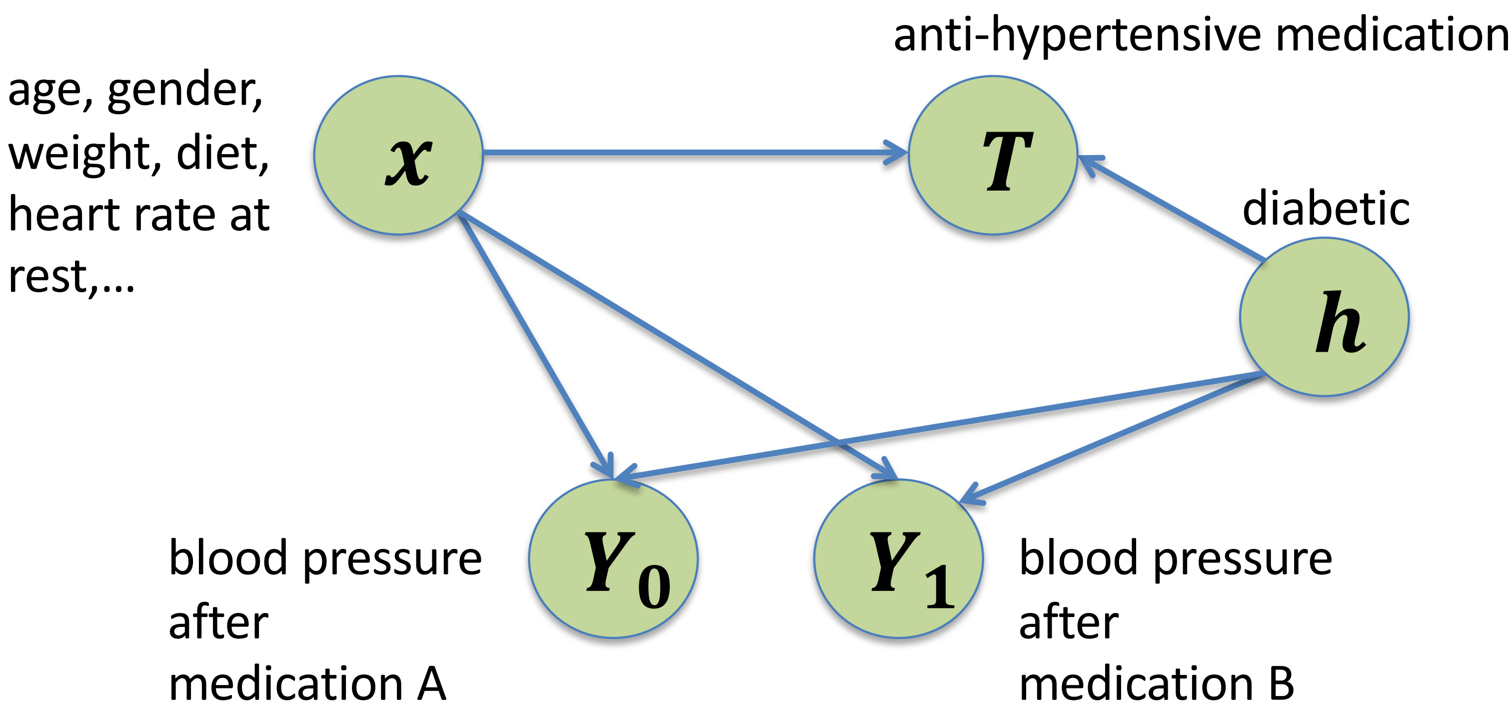
$$(Y_0, Y_1) \perp\!\!\!\perp T \mid x$$

The potential outcomes are independent of treatment assignment, conditioned on covariates x

Typical assumption #2 – no unmeasured confounders



Violation of Typical assumption #2 – ~~no~~ unmeasured confounders



Two common approaches for counterfactual inference

Covariate adjustment

Propensity scores

Covariate adjustment (parametric g-formula)

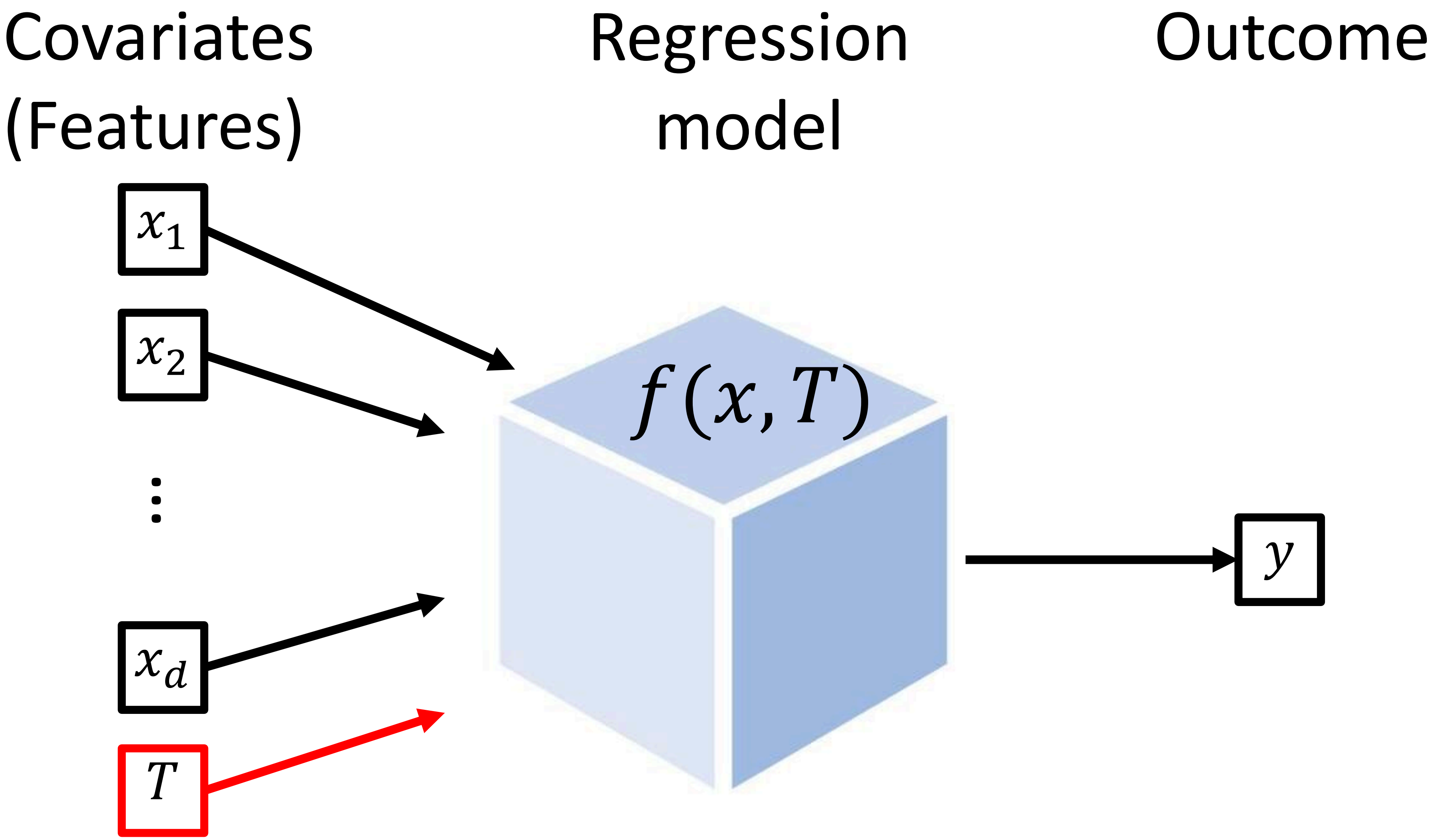
- Explicitly model the relationship between treatment, confounders, and outcome
- We will show that if no unmeasured confounders, expected causal effect of T on Y (given x) is given by:
  $\mathrm{CATE}(\mathrm{x}) = \mathbb{E}[Y|T=1,x] - \mathbb{E}[Y|T=0,x]$
- Fit a model $f(x,t) \approx \mathbb{E}[Y|T=t,x]$

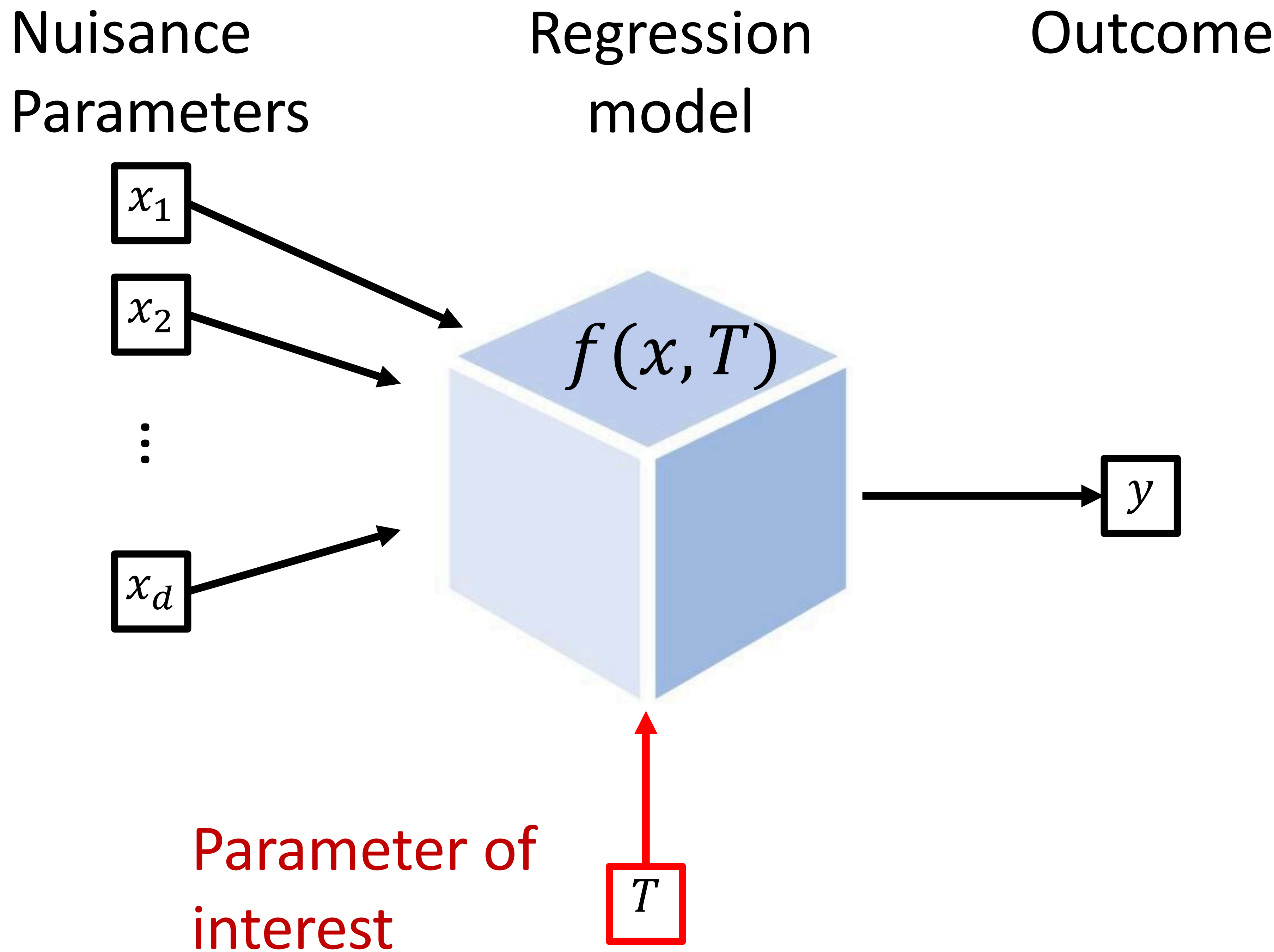
$$\widehat{CATE}(x_i) = f(x_i,1) - f(x_i,0)$$

Covariate adjustment (parametric g-formula)

- Explicitly model the relationship between treatment, confounders, and outcome
- We will show that if no unmeasured confounders, expected causal effect of T on Y (given x) is given by:
  $\text{CATE}(\text{x}) = \mathbb{E}[Y|T=1,x] - \mathbb{E}[Y|T=0,x]$
- Fit a model $f(x,t) \approx \mathbb{E}[Y|T=t,x]$

$$\widehat{ATE} = \frac{1}{n}\sum_{i=1}^{n} f(x_i,1) - f(x_i,0)$$


Covariates
(Features)
Regression
model
Outcome
x_1
x_2
x_d
T
$f(x,T)$
y


Nuisance Parameters
Regression model
Outcome
x_1
x_2
x_d
$f(x, T)$
y
Parameter of interest
T

Average Treatment Effect – the adjustment formula

The expected causal effect of T on Y:

$$ATE := \mathbb{E}\left[Y_1 - Y_0\right]$$

(Hernán & Robins 2010, Pearl 2009)

Average Treatment Effect – the adjustment formula

The expected causal effect of T on Y:

$$ATE := \mathbb{E}\left[Y_1 - Y_0\right]$$

$$\mathbb{E}\left[Y_1\right] =$$

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

(Hernán & Robins 2010, Pearl 2009)

Average Treatment Effect – the adjustment formula

The expected causal effect of T on Y:

$$ATE := \mathbb{E}\left[Y_1 - Y_0\right]$$

$$\mathbb{E}\left[Y_1\right] =$$

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

ignorability

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

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

(Hernán & Robins 2010, Pearl 2009)

Average Treatment Effect – the adjustment formula

The expected causal effect of T on Y:

$$ATE := \mathbb{E}\left[Y_1 - Y_0\right]$$

$$\mathbb{E}\left[Y_1\right] =$$

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

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

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

shorter notation

Average Treatment Effect – the adjustment formula

The expected causal effect of T on Y:

$$ATE := \mathbb{E}\left[Y_1 - Y_0\right]$$

$$\mathbb{E}\left[Y_0\right] =$$

$$\mathbb{E}_{x\sim p(x)}\left[\mathbb{E}_{Y_0\sim p(Y_0|x)}\left[Y_0|x\right]\right] =$$

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

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

Average Treatment Effect – the adjustment formula

Under the assumption of ignorability, we have that:

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

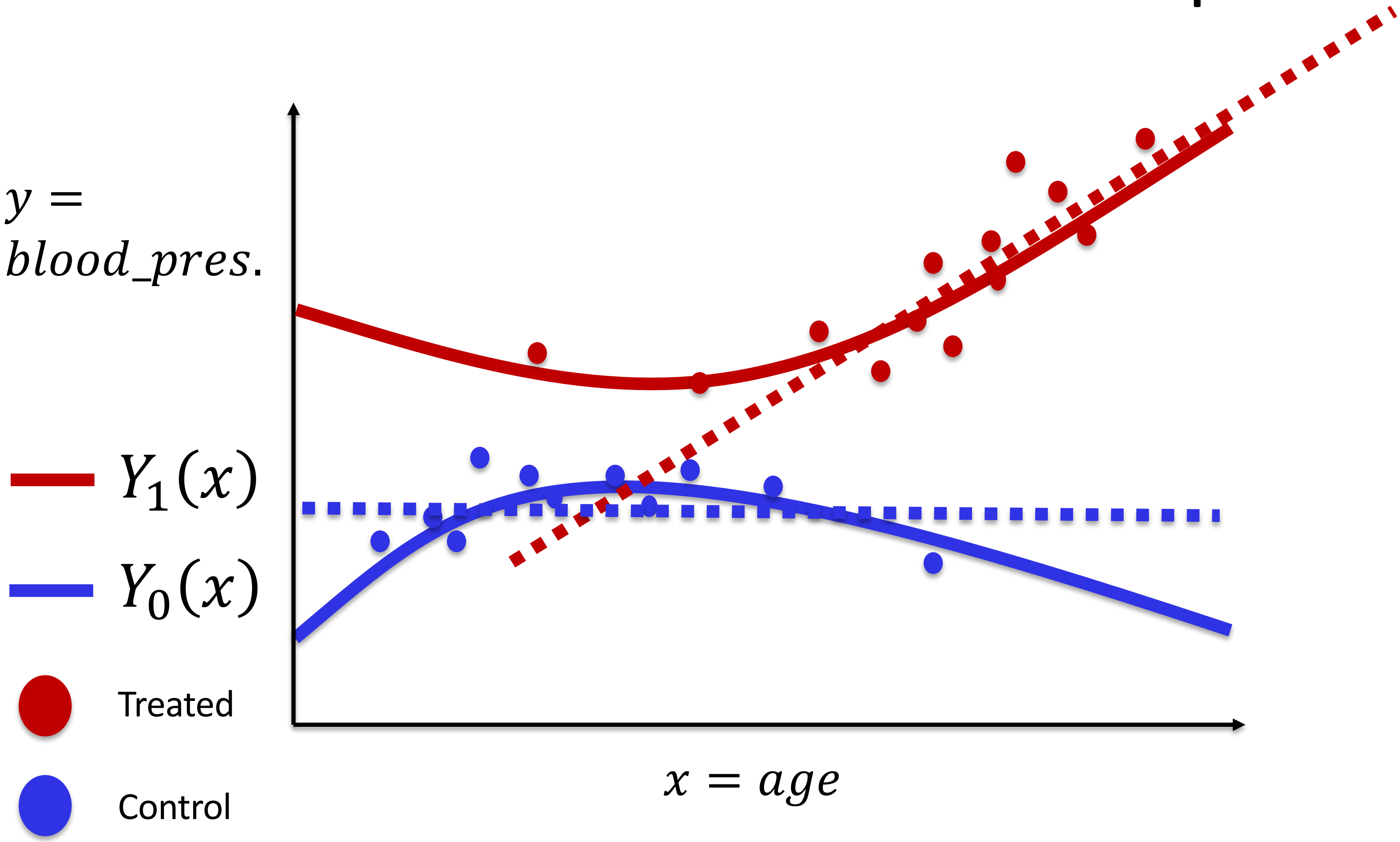
$$\mathbb{E}_{x\sim p(x)}\Big[\, \underbrace{\mathbb{E}\left[Y_1|x, T=1\right] - \mathbb{E}\left[Y_0|x, T=0\right]}\,\Big]$$

$$\left.\begin{array}{l}\mathbb{E}\left[Y_1|x, T=1\right] \\ \mathbb{E}\left[Y_0|x, T=0\right]\end{array}\right\}$$

Quantities we can estimate from data

Empirically we have samples from $p(x|T=1)$ or $p(x|T=0)$. *Extrapolate* to $p(x)$

Example of how covariate adjustment fails when there is no overlap



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:

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

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

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

Covariate adjustment with linear models

- Assume that:

Blood pressure age medication

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

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

- 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:

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

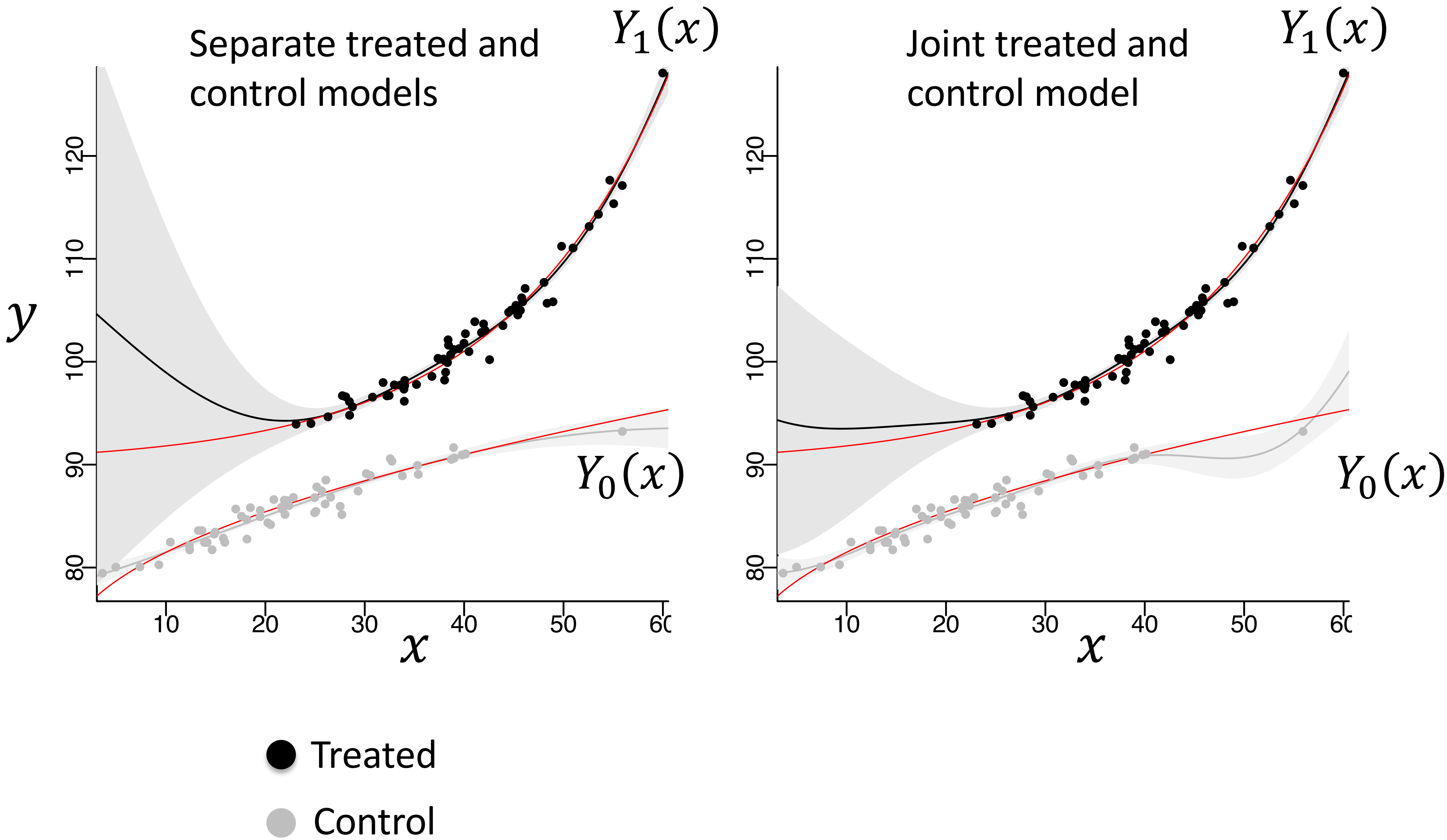
$$\hat{\gamma} = \gamma + \delta \frac{\mathbb{E}[xt]\mathbb{E}[x^2] - \mathbb{E}[t^2]\mathbb{E}[x^2 t]}{\mathbb{E}[xt]^2 - \mathbb{E}[x^2]\mathbb{E}[t^2]}$$

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

Covariate adjustment with non-linear models

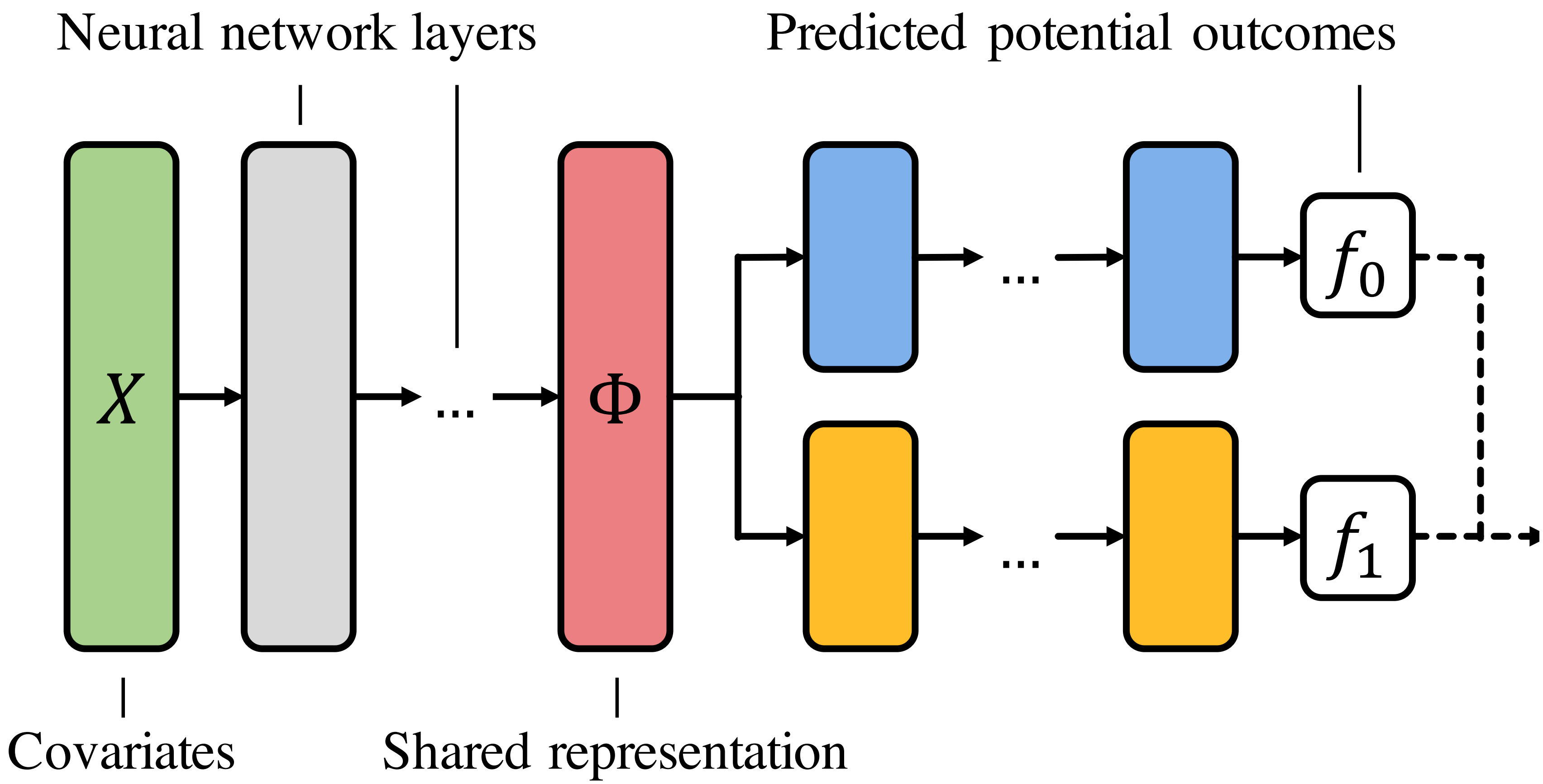
- Random forests and Bayesian trees
  Hill (2011), Athey & Imbens (2015), Wager & Athey (2015)
- Gaussian processes
  Hoyer et al. (2009), Zigler et al. (2012)
- Neural networks
  Beck et al. (2000), Johansson et al. (2016), Shalit et al. (2016), Lopez-Paz et al. (2016)

Example: Gaussian processes



Figures: Vincent Dorie & Jennifer Hill

Example: Neural networks



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

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

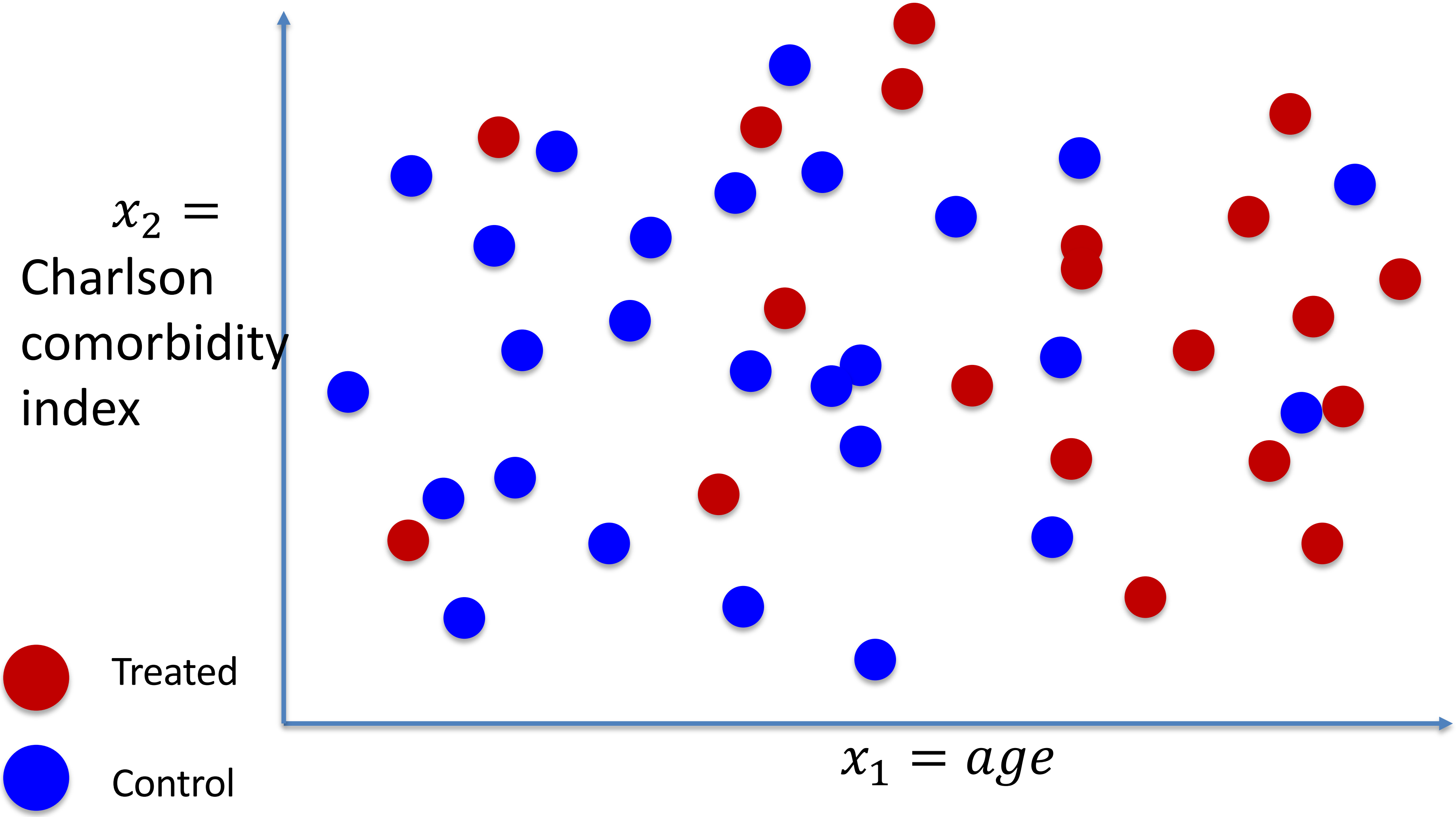
Inverse propensity score re-weighting

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

control *treated*

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

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

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

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

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

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

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

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

Propensity scores - 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

Bounding counterfactual risk

- Building on ML literature from domain adaptation, we can **bound** the (average) error in predicting counterfactuals:

$$\mathbb{E}_{p^{t=0}(x)}\left[\left(Y_1 - f(x,1)\right)^2\right] \leq \mathbb{E}_{p^{t=1}(x)}\left[\left(Y_1 - f(x,1)\right)^2\right] + |\ell_f|_{\mathcal{H}} d_{\mathcal{H}}\left(p^{t=0}(x), p^{t=1}(x)\right)$$

Counterfactual risk **Factual risk** **Distance between treatment groups**

- Makes no assumption of **consistency or overlap**
- Suggests avenues for modifying empirical risk minimization when used for counterfactual inference

Johansson, Shalit, S. *ICML*. 2016; Shalit, Johansson, S. *ICML*. 2017

Bounding counterfactual risk

- Building on ML literature from domain adaptation, we can **bound** the (average) error in predicting counterfactuals:

$$\mathbb{E}_{p^{t=0}(x)}\left[\left(Y_1 - f(x,1)\right)^2\right] \leq \mathbb{E}_{p^{t=1}(x)}\left[w_1(x)\left(Y_1 - f(x,1)\right)^2\right] + |\ell_f|_{\mathcal{H}}\, d_{\mathcal{H}}\left(p^{t=0}(x), w_1(x)p^{t=1}(x)\right)$$

Counterfactual risk **Factual risk** **Distance between treatment groups**

- Makes no assumption of **consistency** or **overlap**
- For example, here we minimize an importance weighted empirical risk minimization, where weights can be learned

Johansson, Kallus, Shalit, Sontag, 2018

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
- Consistency of estimates depend on:
    - Causal graph being correct (e.g., no unobserved confounding)
    - Identifiability of causal effect (e.g., overlap)
    - Correctly specified models

References

- Also discussed in class: Instrumental variables, doubly robust estimators
- 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”)
- A few recent papers touching on topics we discussed in class:
  https://arxiv.org/abs/1906.02120
  https://arxiv.org/abs/1705.08821
  https://arxiv.org/abs/1510.04342
  https://arxiv.org/abs/1810.02894