Topics in Machine Learning
Machine Learning for Healthcare

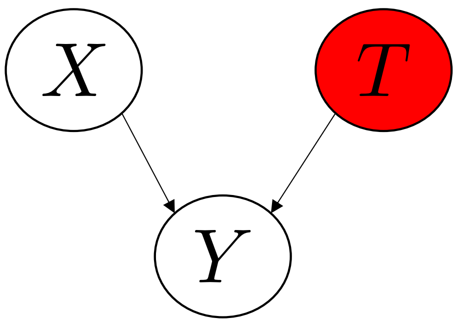
Rahul G. Krishnan
Assistant Professor
Computer science & Laboratory Medicine and Pathobiology

Slide credits to David Sontag & Uri Shalit

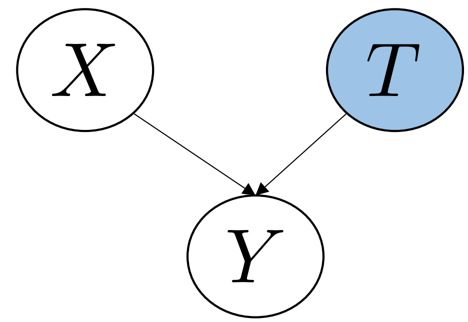
Outline – A taste of causal inference

- Recap: Randomized control trials
- Making decisions with predictive models
- Potential outcomes
- Do operator
- Assumptions in causal inference
- Algorithms for estimating ATE and CATE

Last week



Treatment

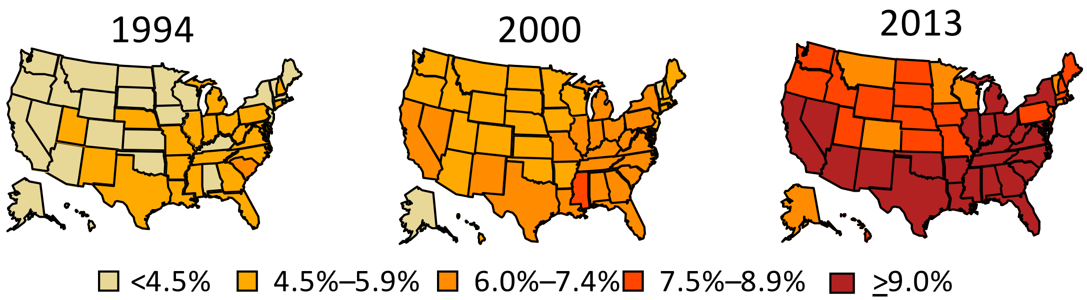


Control

$$ATE = \sum_{i \in \mathcal{T}} y_t - \sum_{i \in \mathcal{C}} y_c \approx \mathbb{E}[Y_t - Y_c]$$

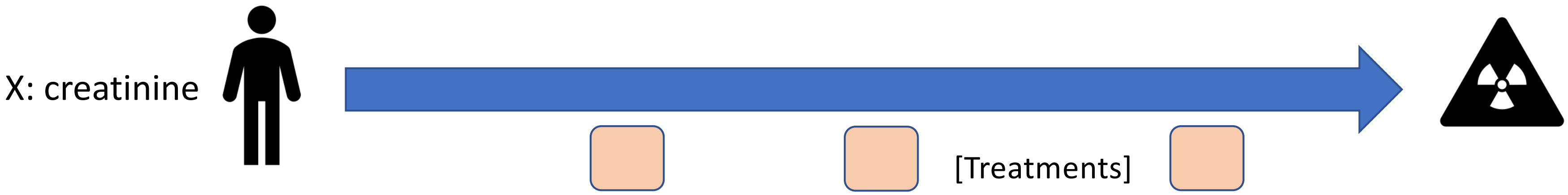
Y is an outcome of interest (positive value good; negative value bad)

Risk stratification



- We studied how supervised machine learning could be used for the early detection of diabetes
- There were several features that were associated with an increased/decreased risk of diabetic onset
- Gastric bypass was the highest-negative weight
  - **Is this a good idea for an intervention?**

Survival analysis

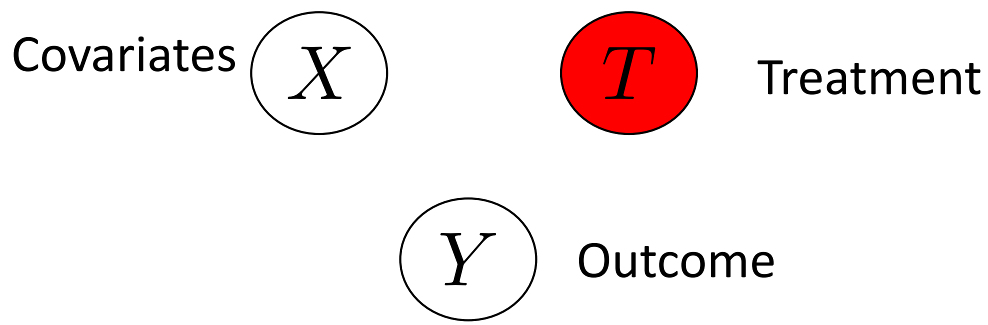


- We can use survival models to predict time of dying/progression of disease in the future
- Trained a CoxPH model and feature associated with high creatinine increased survival time
  - Should we conclude that increasing creatinine improved survival?

Challenges in making decisions from healthcare data

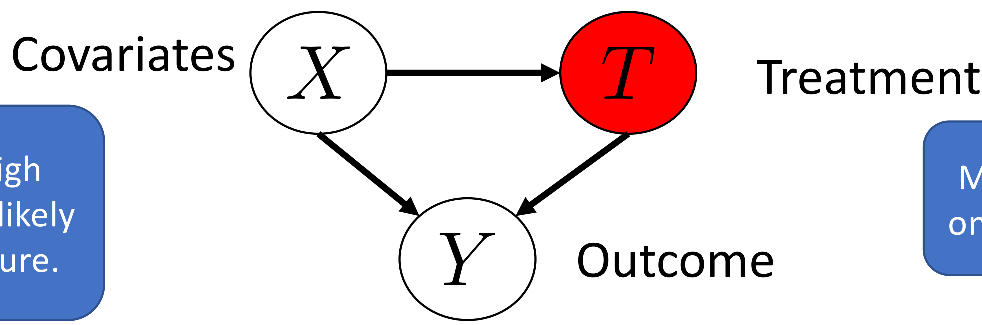
- Doing an RCT is not always ethical
  - e.g. cannot force a control group to smoke to assess its effects on long-cancer
- Can we make causal conclusions from *retrospective* data?
- Key challenges:
  - Unobserved confounding
  - Positive support

What should the graph of observational health data look like?



A simple causal graph for observational data

Doctor might prescribe Metformin if A1C level is high



Patient could have a high BMI making them more likely to suffer from heart failure.

Metformin has an effect on an outcome like death

Causal inference with the Potential Outcomes framework

- Each individual x_i has two potential outcomes:
  - Control outcome [PO had the individual not been treated] $Y_0(x_i)$
  - Treated outcome [PO had the individual been treated] $Y_1(x_i)$
- Conditional average treatment effect [CATE]

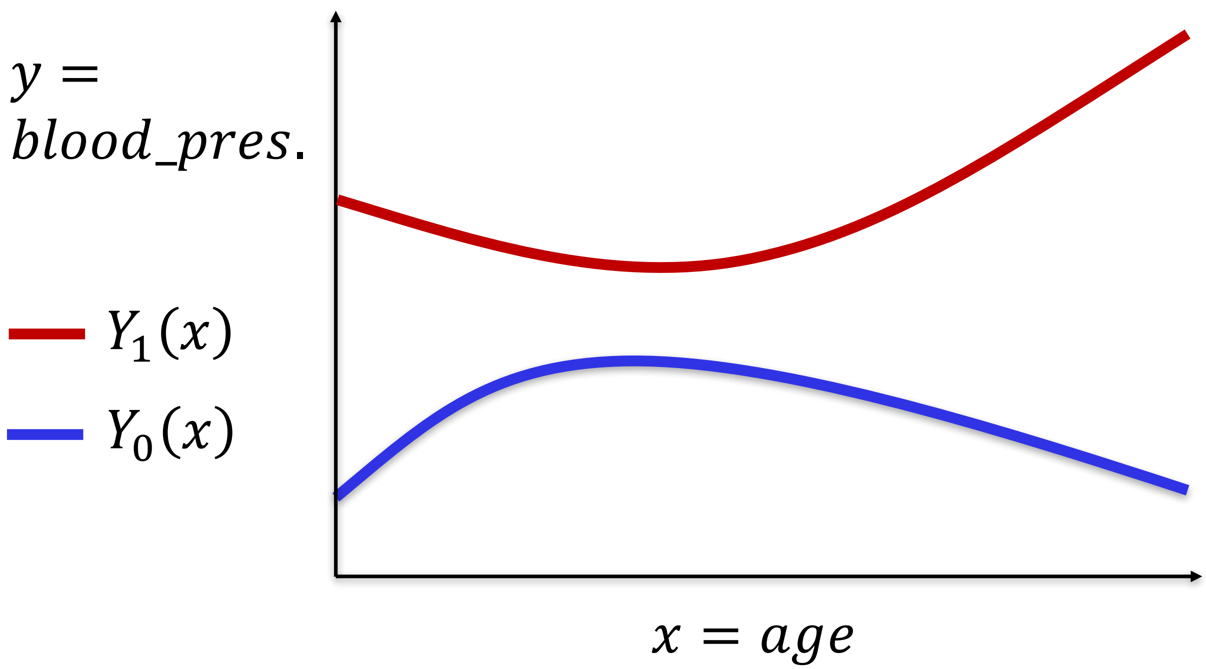
$$\mathrm{CATE}(x_i) = \mathbb{E}_{Y_1 \sim p(Y_1|x_i)}[Y_1|x_i] - \mathbb{E}_{Y_0 \sim p(Y_0|x_i)}[Y_0|x_i]$$

- Average treatment effect [ATE]

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

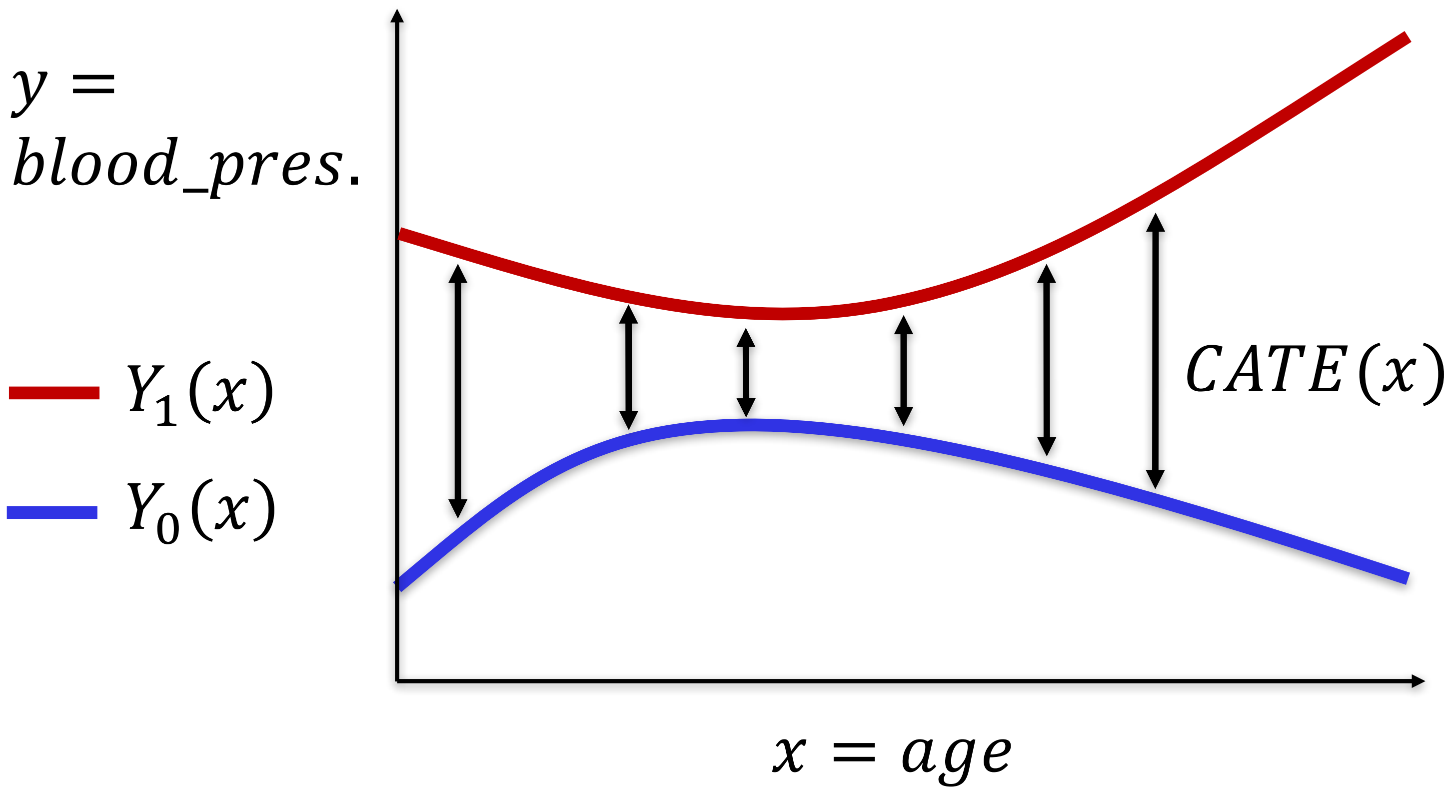
Visualizing notation on potential outcomes

Example – Blood pressure and age

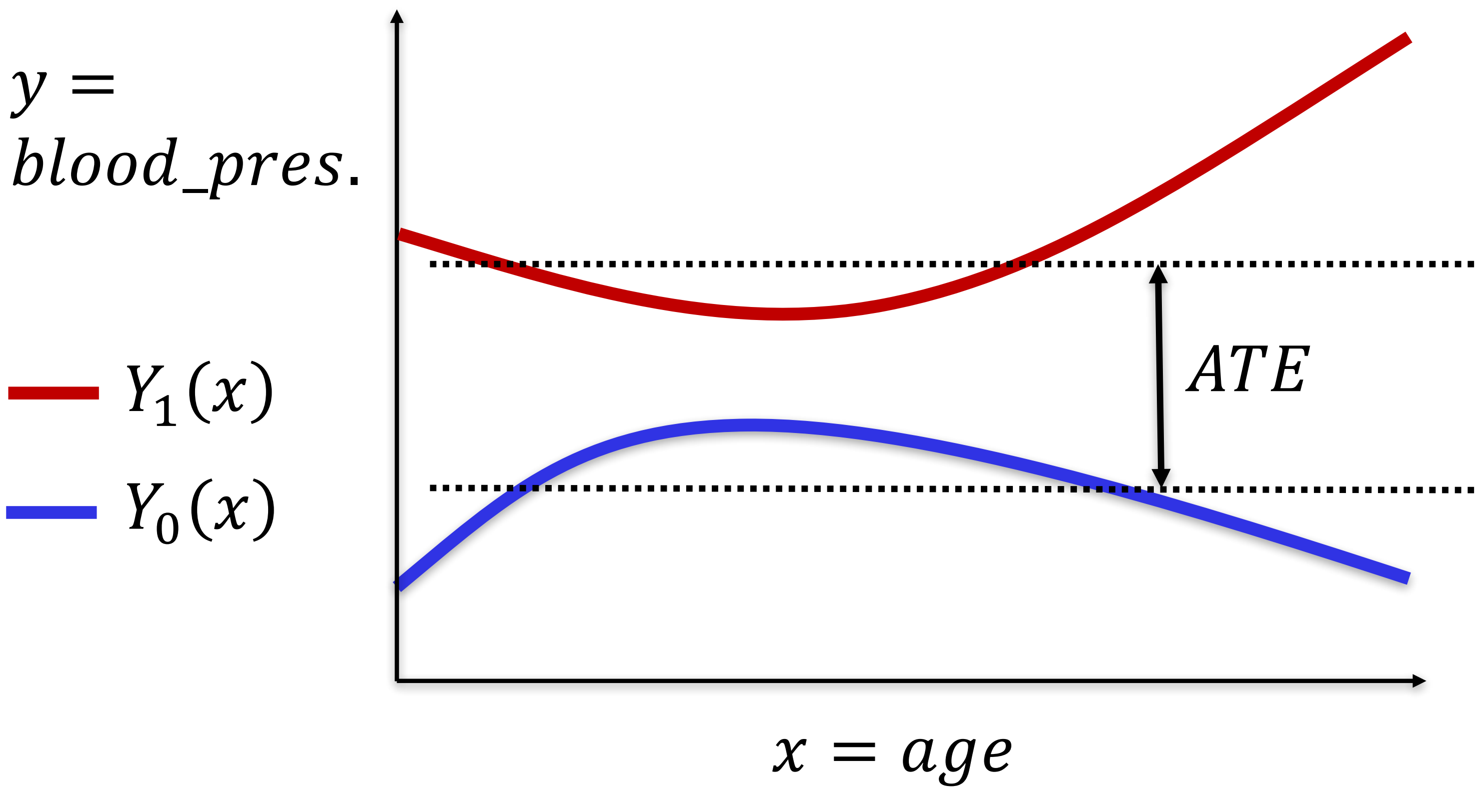


Slide credit: David Sontag

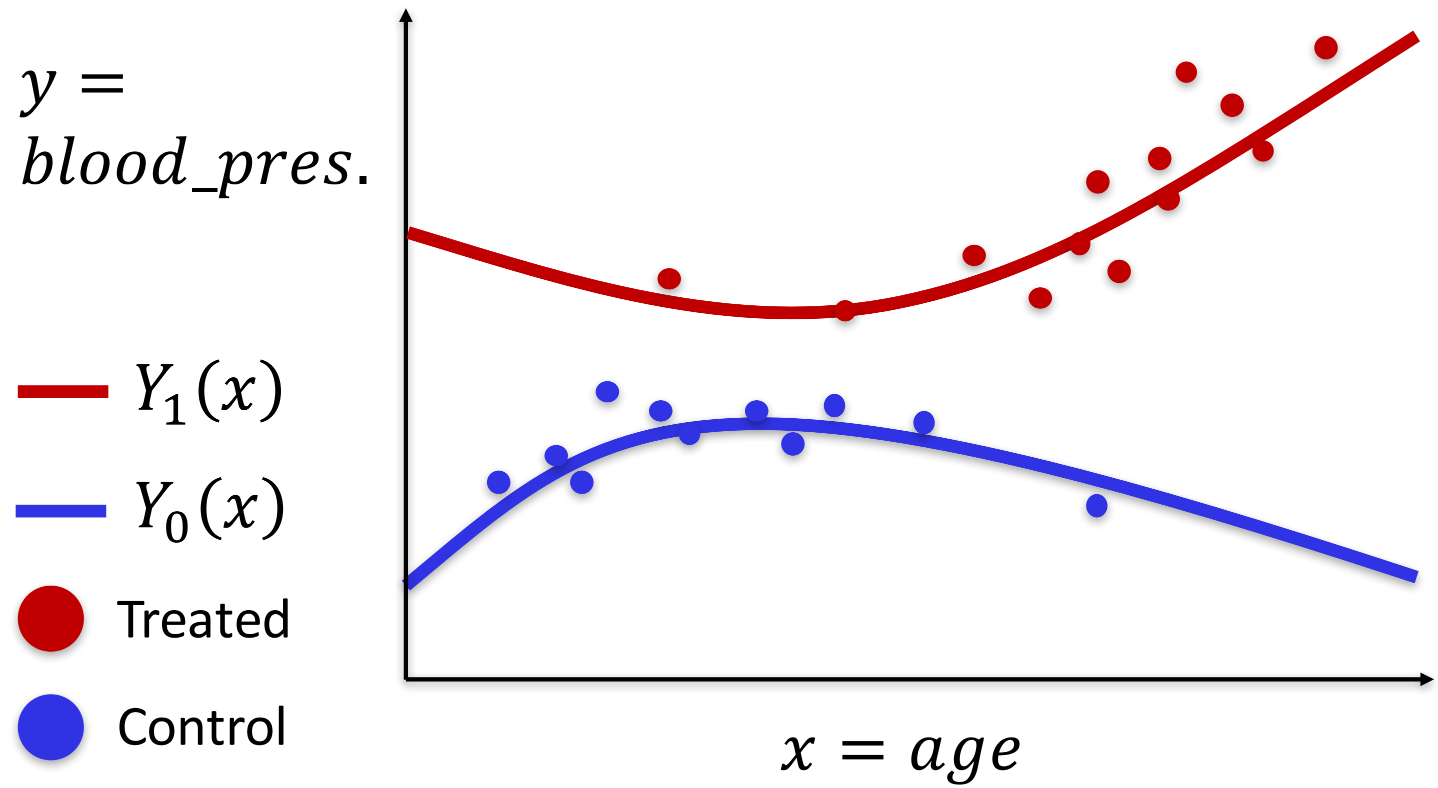
Blood pressure and age



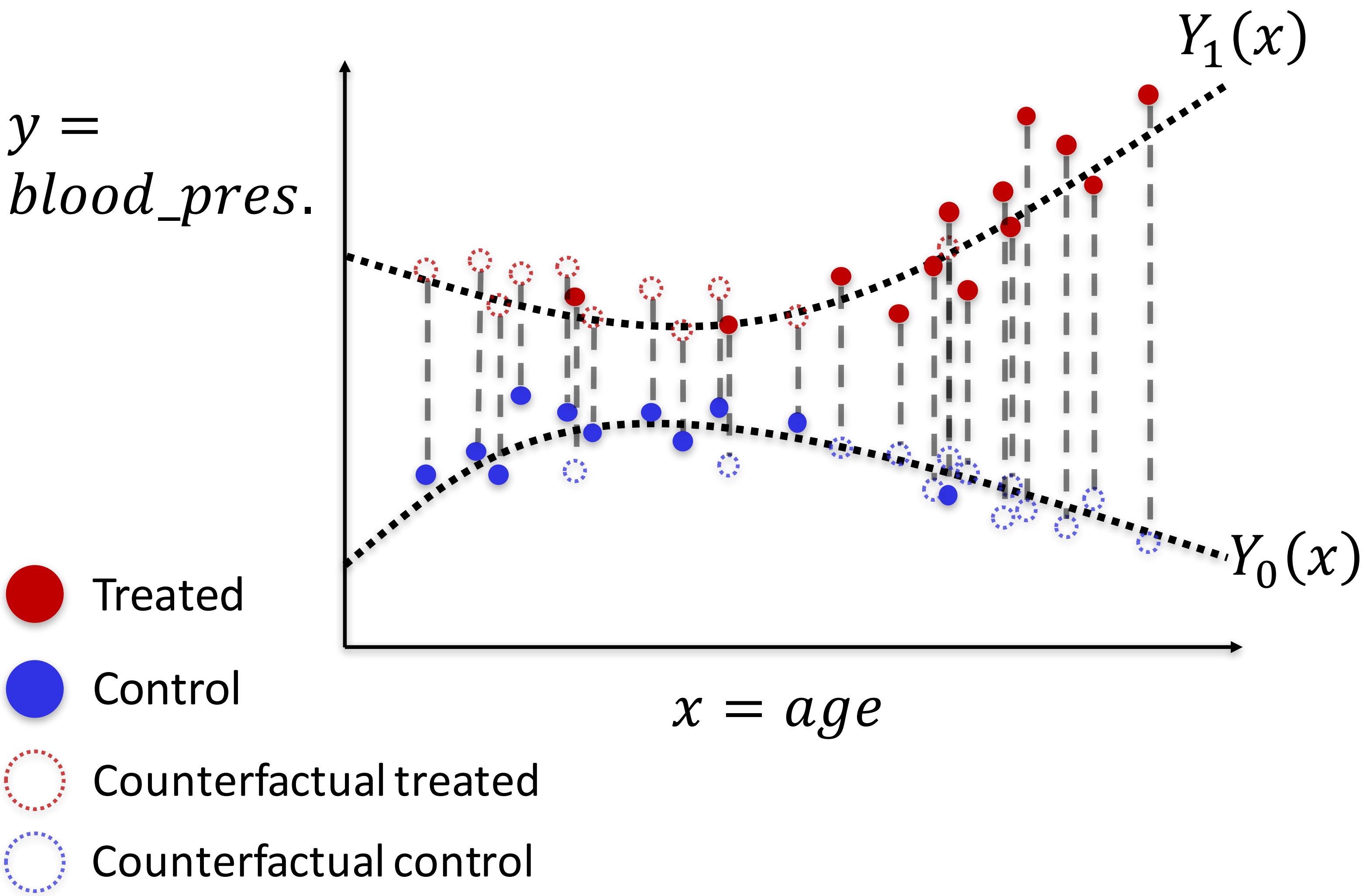
Blood pressure and age



Blood pressure and age



Blood pressure and age



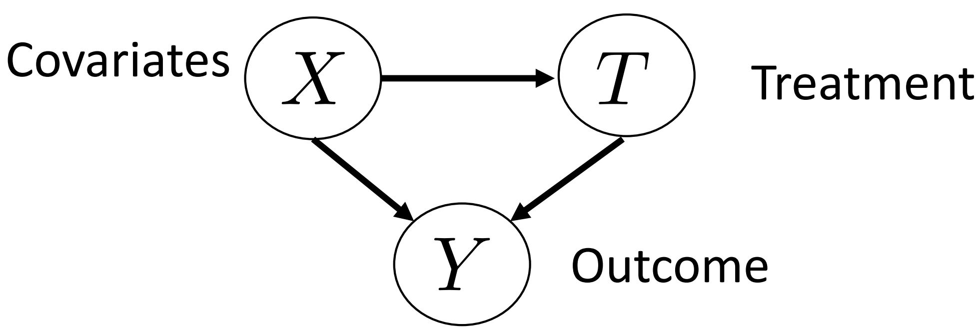
(age, gender, exercise)	Y_0: Sugar levels *had they received* medication A	Y_1: Sugar levels *had they received* medication B	Observed sugar levels
(45, F, 0)	**6**	5.5	6
(45, F, 1)	7	**6.5**	6.5
(55, M, 0)	**7**	6	7
(55, M, 1)	9	**8**	8
(65, F, 0)	8.5	**8**	8
(65,F, 1)	**7.5**	7	7.5
(75,M, 0)	10	**9**	9
(75,M, 1)	**8**	7	8

(Example from Uri Shalit)

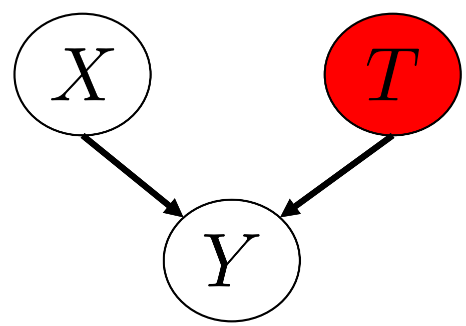
For an individual we only observe one of the potential outcomes

- *Fundamental problem of causal inference* (Rubin 1974; Holland 1986)
- ~~*How*~~*When* can we make causal conclusions despite this?

Learning from retrospective data

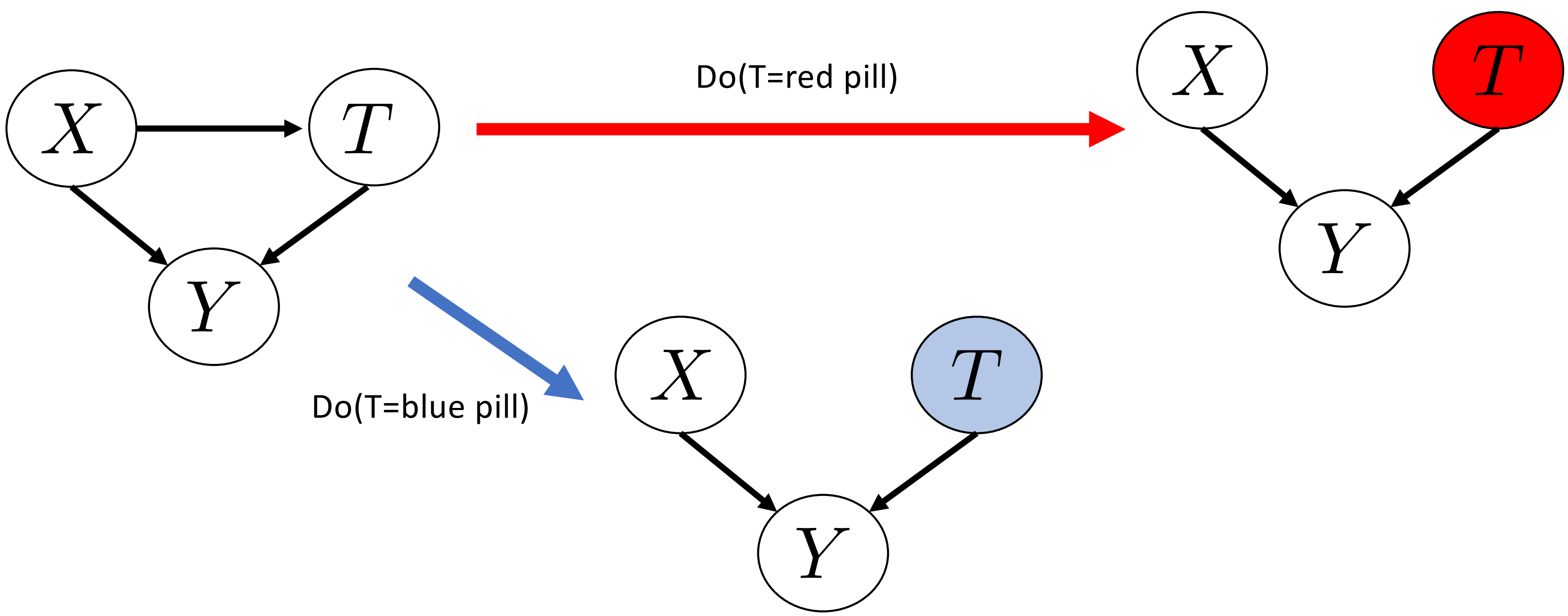


Learning from interventional data [RCTs]



The do-operator

- The do-operator (Pearl, 2000) is a graphical operator on a causal graph that characterizes the effect of an intervention
- It allows us to mimic the effect of an intervention onto variation in the joint probability distribution



Causal inference

- **What we have:** Data is drawn from the joint distribution on the left
- **What we want:** Samples from the joint distributions on the right
- **Key idea:** Under certain assumptions, we can estimate conditional distributions from the right
- **Strategy:** Write down the causal estimand using quantities estimable from the observed (left) distribution)

Assumptions in causal inference – (1)

Y_0, Y_1: potential outcomes for control and treated

x: unit covariates (features)

T: treatment assignment

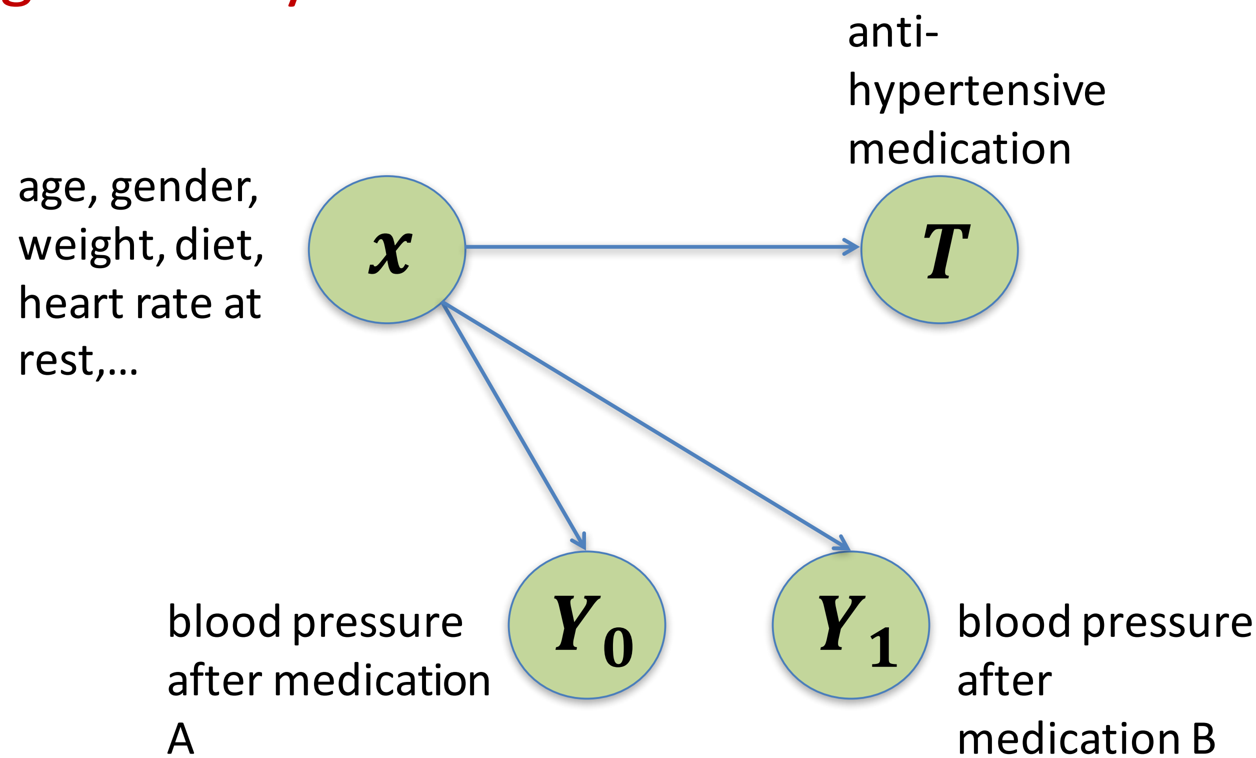
We assume:

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

Also known as Ignorability

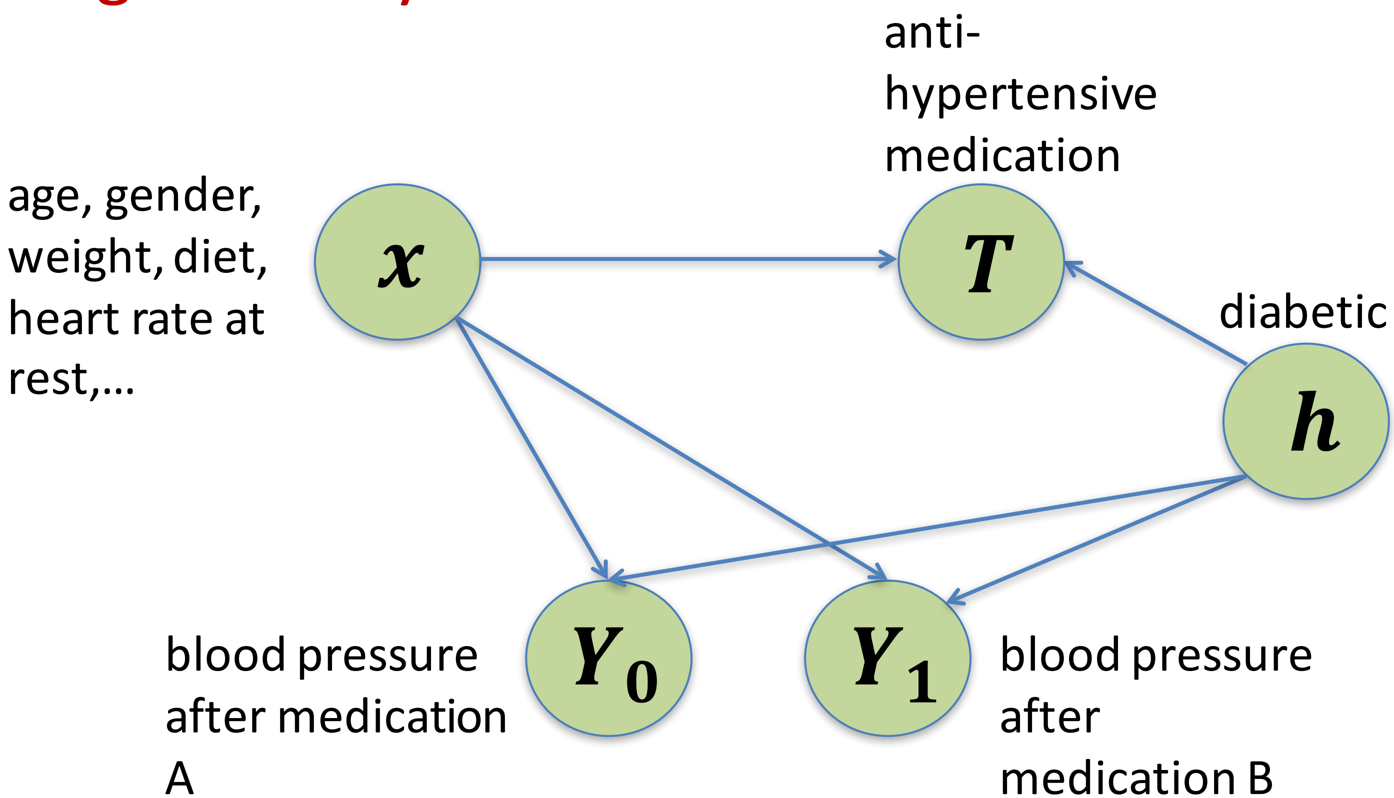
The potential outcomes are independent of treatment assignment, conditioned on covariates x

Ignorability



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

No Ignorability



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

Assumptions in causal inference – (2)

Y_0, Y_1: potential outcomes for control and treated

x: unit covariates (features)

T: treatment assignment

We assume:

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

Also known as Common Support

Before beginning any causal analysis

- Understand where the data is coming from (are there biases you didn’t account for)
- Drawing the causal graph (or even approximations to it) can be a crucial exercise
- What should X, T, Y be? Do they satisfy ignorability and positivity?
- What is the causal query of interest?

Adjustment formula (Hernan and Robins, 2010) (Pearl, 2009)

- Also known as the G-formula, it provides a mechanism to write down a causal estimand using observational data

Average Treatment Effect

The expected causal effect of T on Y:

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

Average Treatment Effect

The expected causal effect of T on Y:

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

law of total expectation

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

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

Average Treatment Effect

The expected causal effect of T on Y:

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

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

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

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

ignorability

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

Average Treatment Effect

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 expected causal effect of T on Y:

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

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

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

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

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

The adjustment formula

Under the assumption of ignorability, we have that:

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

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

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

Quantities we can estimate from data

The adjustment formula

Under the assumption of ignorability, we have that:

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

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

$$\mathbb{E}\left[Y_0 | x, T=1\right]$$

$$\mathbb{E}\left[Y_1 | x, T=0\right]$$

$$\mathbb{E}\left[Y_0 | x\right]$$

$$\mathbb{E}\left[Y_1 | x\right]$$

Quantities we *cannot* directly estimate from data

The adjustment formula

Under the assumption of ignorability, we have that:

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

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

$\mathbb{E}\left[Y_1|x, T=1\right]$

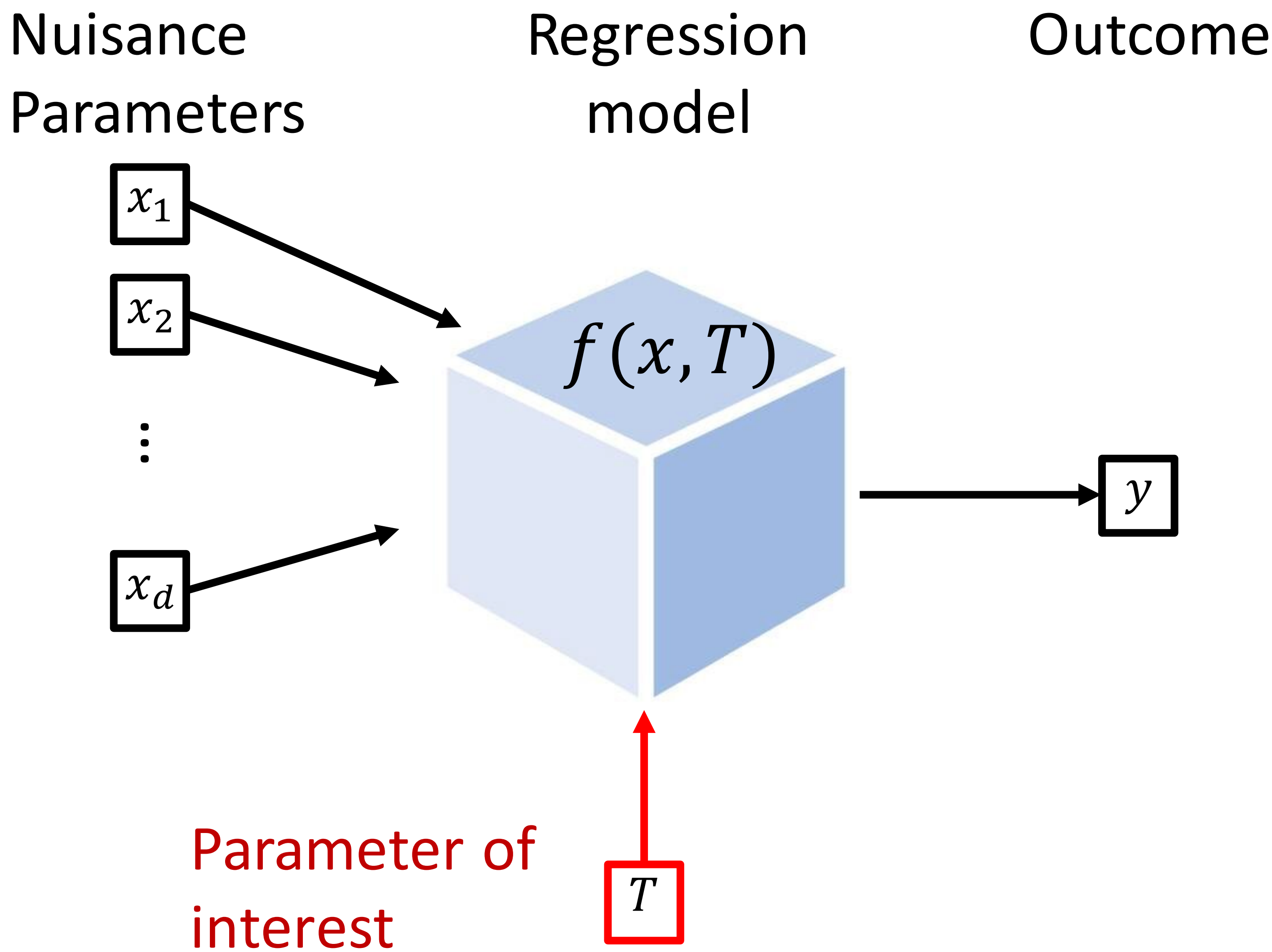
$\mathbb{E}\left[Y_0|x, T=0\right]$

Quantities we can estimate from data

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

Strategies for Adjustment

- Covariate Adjustment
  - Response surface modeling
  - Use a parametric model of treatments, confounders and outcome



Covariate adjustment (parametric g-formula)

- Explicitly model the relationship between treatment, confounders, and outcome
- Under ignorability, the expected causal effect of T on Y:
  $\mathbb{E}_{x \sim p(x)}\left[\, \mathbb{E}[Y_1 | T = 1, x] - \mathbb{E}[Y_0 | T = 0, x] \right]$
- Fit a model $f(x,t) \approx \mathbb{E}[Y_t | T = t, x]$

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

Covariate adjustment (parametric g-formula)

- Explicitly model the relationship between treatment, confounders, and outcome
- Under ignorability, the expected causal effect of T on Y:
  $\mathbb{E}_{x\sim p(x)}\big[\,\mathbb{E}[Y_1|T=1,x]-\mathbb{E}[Y_0|T=0,x]\big]$
- Fit a model $f(x,t)\approx\mathbb{E}[Y_t|T=t,x]$

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

Covariate adjustment with linear models

- Assume that:

Blood pressure age medication

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

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

- 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!

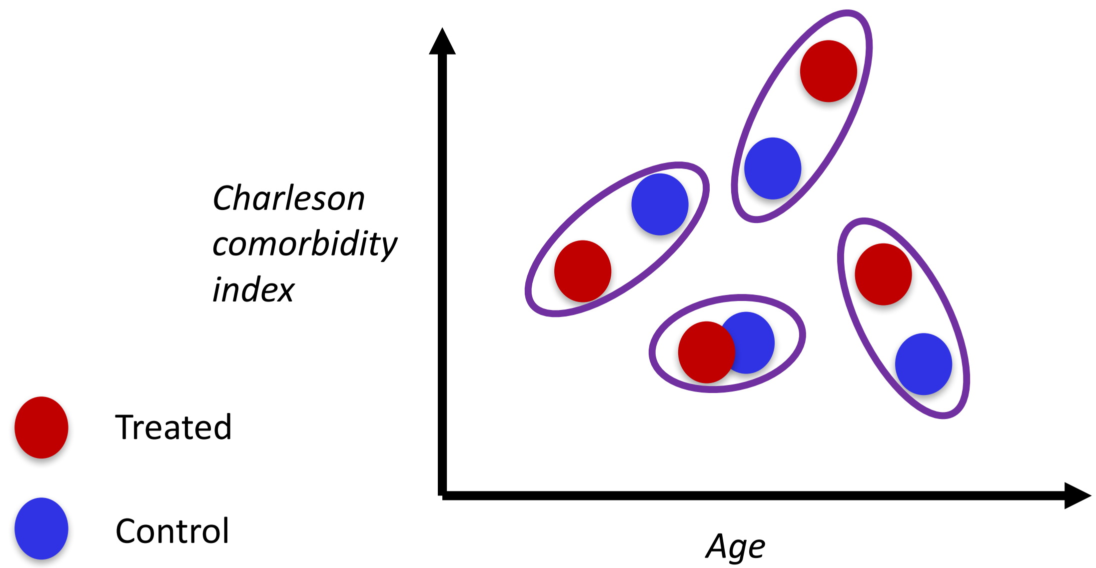
Covariate adjustment with non-linear models

- 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)

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)

Alternative strategies for adjustment - Matching



Summary

- Two strategies where machine learning may be used for causal inference:
  - Predict outcome given features and impute counterfactuals [covariate adjustment]
  - Predict treatment using features (propensity scores) and use to reweigh the outcome

Causal inference is a big field

- Causal inference has been studied in many communities of science including economics, statistics, machine learning
- Different schools of thought – causal graphs vs conditional independence statements
- Beyond the scope to have an in-depth discussion of the all techniques underlying this field in this class
- Lots of research in leveraging ideas from causal inference to improve predictive models