MACHINE LEARNING FOR HEALTHCARE

6.S897, HST.S53

Lecture 3: Causal inference

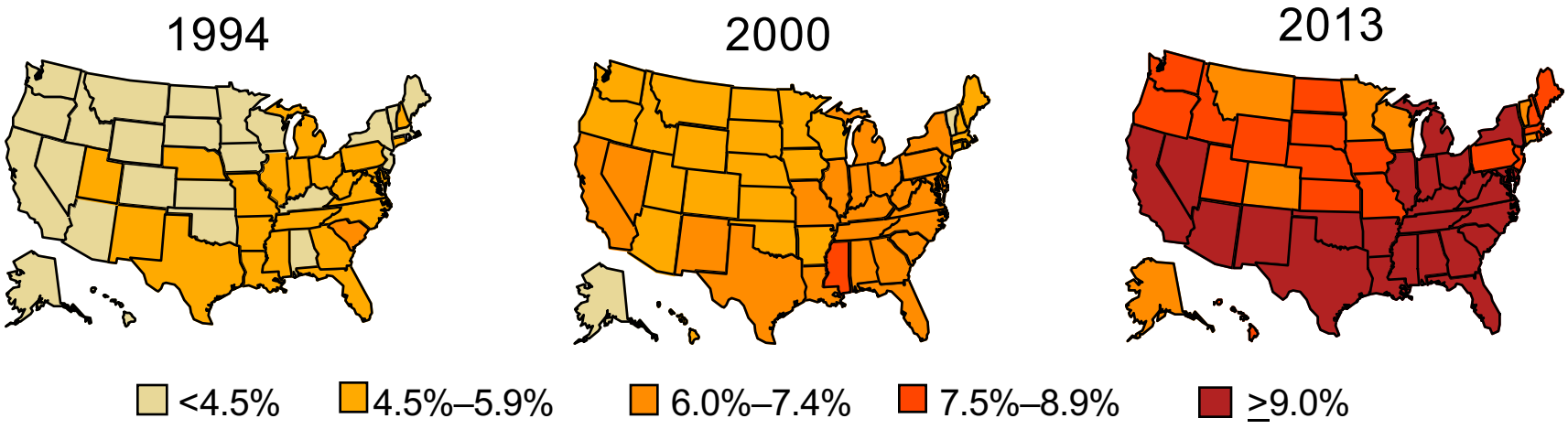
Prof. David Sontag

MIT EECS, CSAIL, IMES

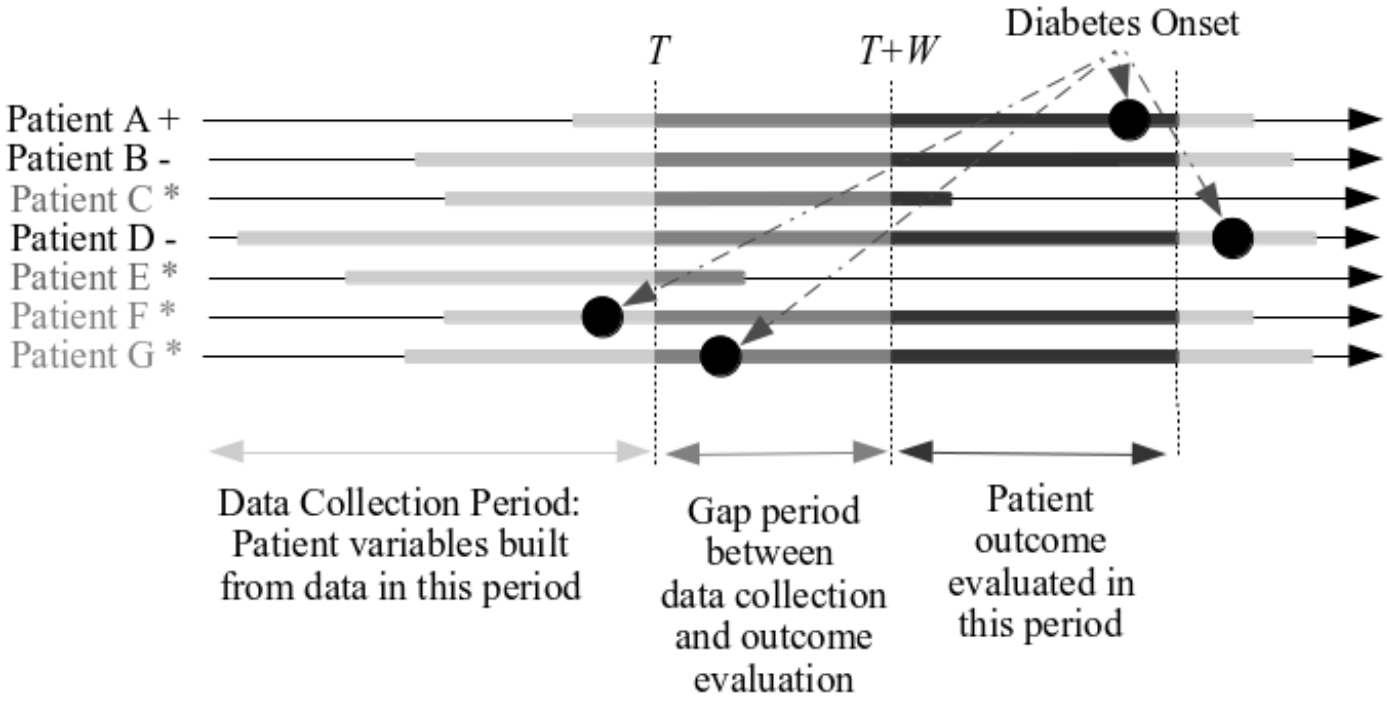
(Thanks to Uri Shalit for many of the slides)

Massachusetts Institute of Technology

*Last week: Type 2 diabetes



Early detection of Type 2 diabetes:



(Razavian et al., *Big Data*, 2016)

*Last week: Discovered risk factors

Highly weighted features	O
Impaired Fasting Glucose (Code 790.21)	
Abnormal Glucose NEC (790.29)	
Hypertension (401)	
Obstructive Sleep Apnea (327.23)	
Obesity (278)	
Abnormal Blood Chemistry (790.6)	
Hyperlipidemia (272.4)	
Shortness Of Breath (786.05)	
Esophageal Reflux (530.81)	

Additional Disease Risk Factors Include:
Pituitary dwarfism (253.3), Hepatomegaly(789.1), Chronic Hepatitis C (070.54), Hepatitis (573.3), Calcaneal Spur(726.73), Thyrotoxicosis without mention of goiter(242.90), Sinoatrial Node dysfunction(427.81), Acute frontal sinusitis (461.1), Hypertrophic and atrophic conditions of skin(701.9), Irregular menstruation(626.4), …

(1.78 1.93)

Diabetes
1-year gap

(Razavian et al., *Big Data*, 2016)

Thinking about interventions

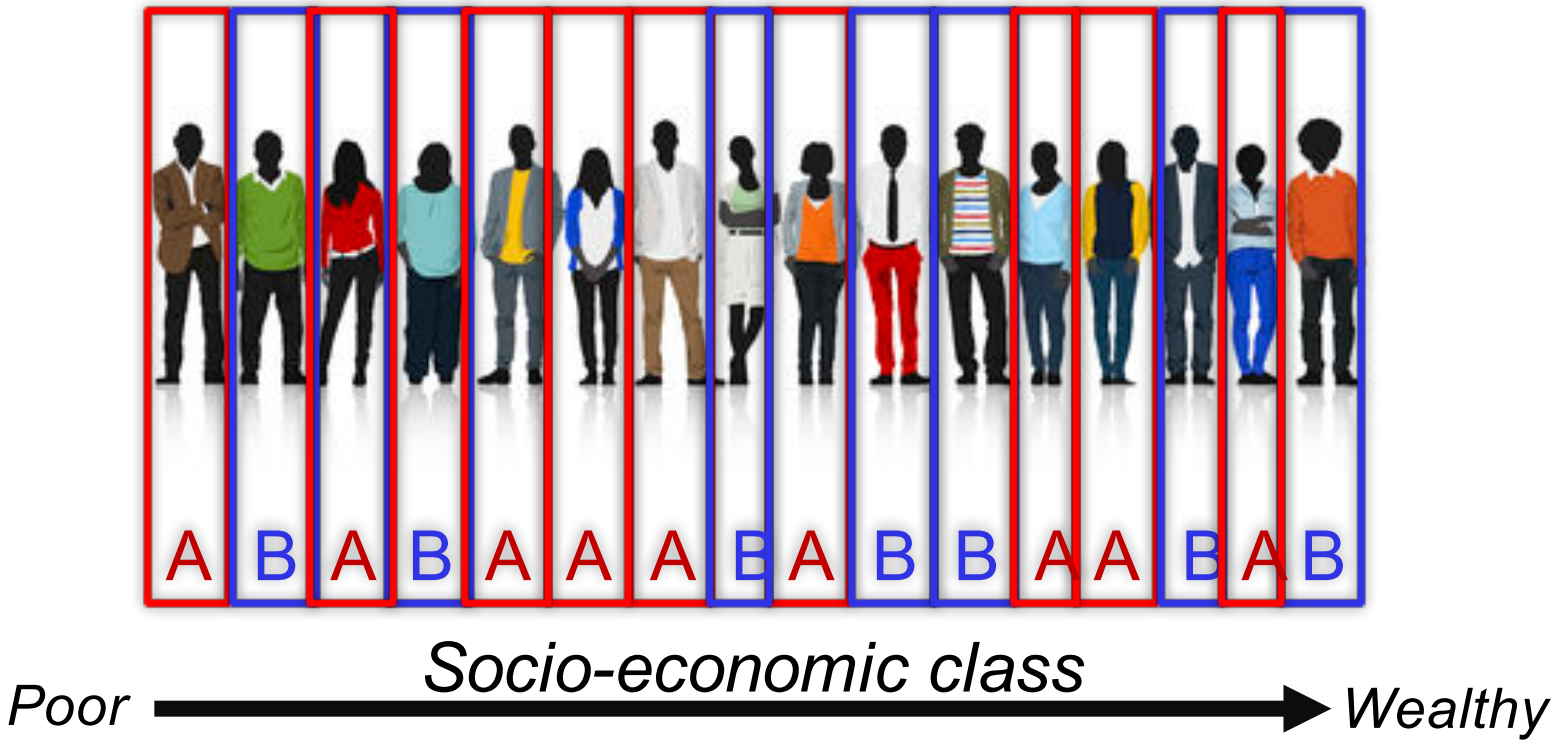
1. Do highly weighted features suggest avenues for *preventing* onset of diabetes?
   - **Example: Gastric bypass surgery**. Highest negative weight (9th most predictive feature)
   - What is the mathematical justification for thinking of highly weighted features in this way?
2. What happens if the patient did *not* get diabetes because an intervention made in the gap?
   - How do we **deconvolve** effect of interventions from the prediction task?
3. Solution is to reframe as causal inference problem: **predict for which patients an intervention will reduce chances of getting T2D**

Randomized trials vs. observational studies

Which treatment works better?

A **or** B

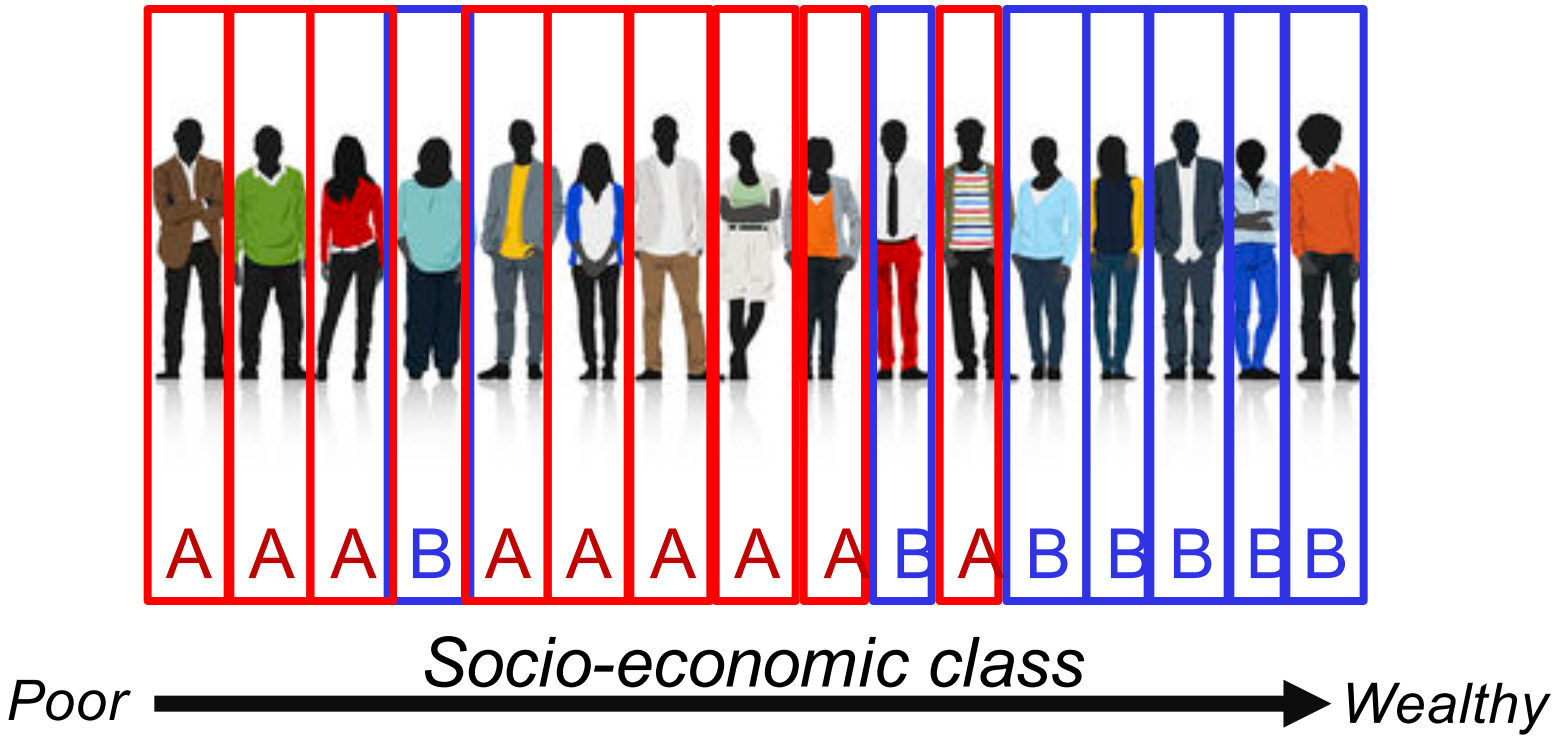
Randomized controlled trial (RCT)



Which treatment works better?

A **or** B

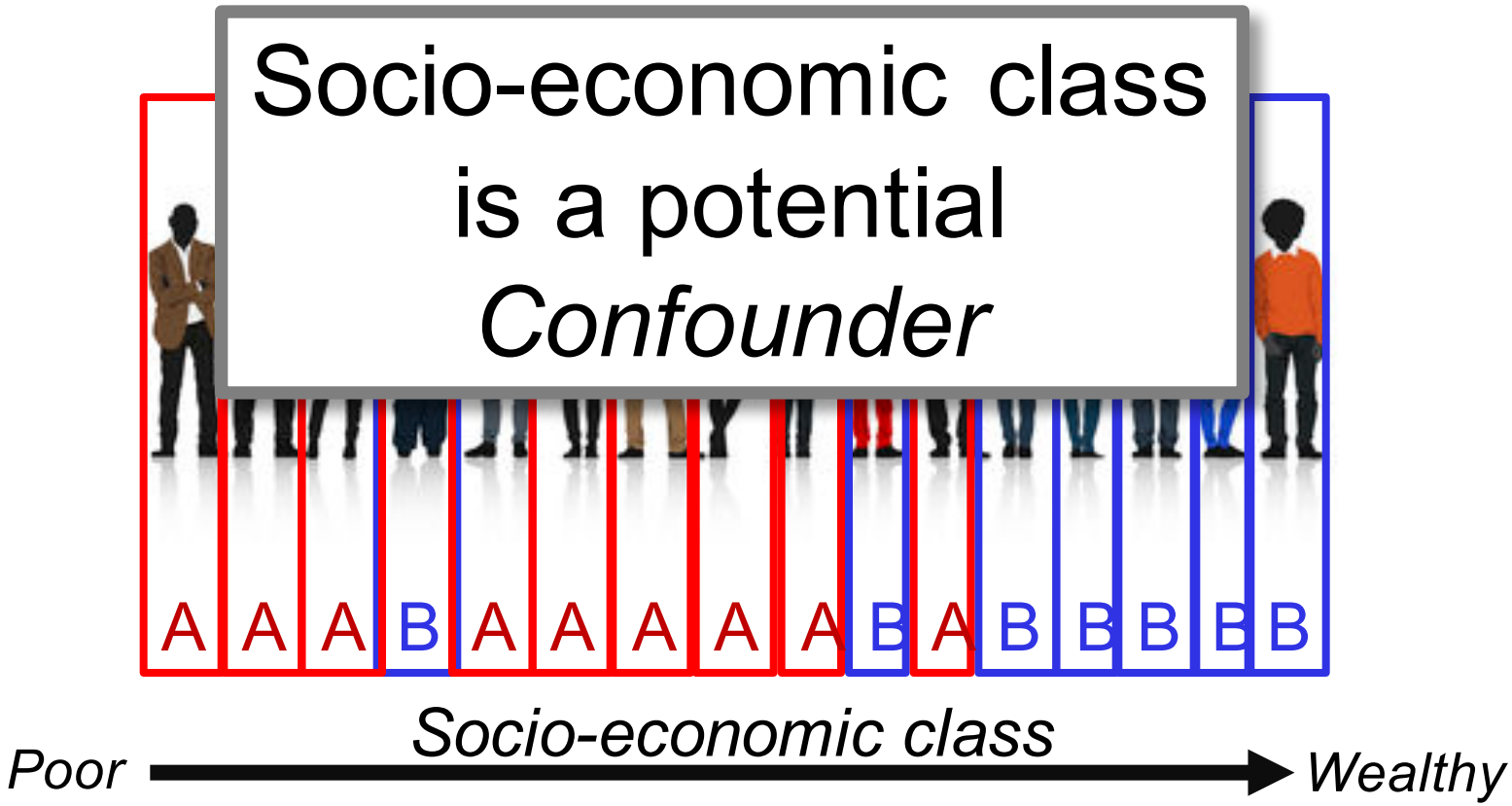
Observational study



Which treatment works better?

A **or** B

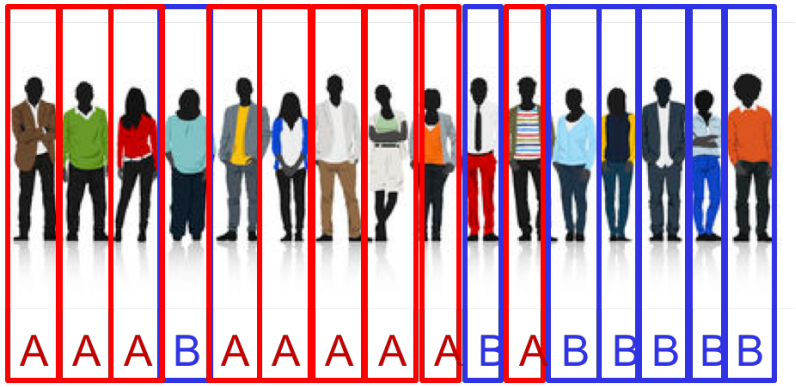
Observational study

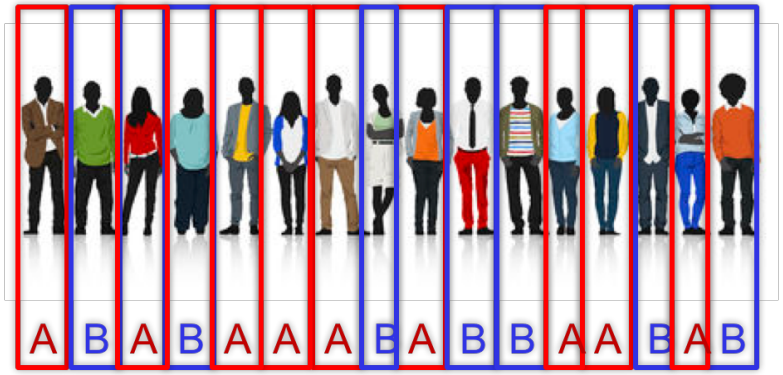


Which treatment works better?

A **or** B

In many fields randomized studies are the gold standard for causal inference, but…



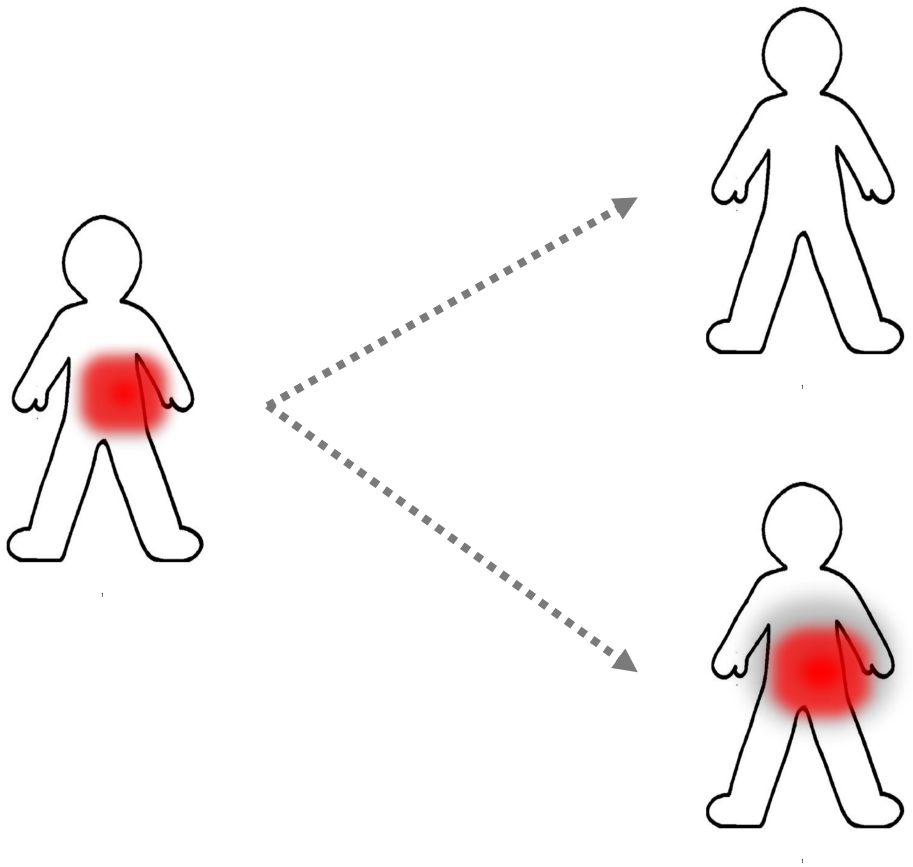


- Does inhaling Asbestos cause cancer?
- Does decreasing the interest rate reinvigorate the economy?
- We have a budget for **one new** anti-diabetic drug experiment. Can we use past health records of 100,000 diabetics to guide us?

Even randomized controlled trials have flaws

- Not personalized – only population effect
- Study population might not represent true population
  - Recruiting is hard
  - People might drop out of study
  - Study in one company/hospital/state/country could fail to generalize to others

Example 1
Precision medicine:
Individualized Treatment Effect (ITE)



Which treatment is best **for me**?

- Which anti-hypertensive treatment?
  - Calcium channel blocker (A)
  - ACE inhibitor (B)
- Current situation:
  - Clinical trials
  - Doctor's knowledge & intuition
- Use datasets of patients and their histories

- *Blood pressure = 150/95*
- *WBC count = 6*10^9/L*
- *Temperature = 98°F*
- *HbA1c = 6.6%*
- *Thickness of heart artery plaque = 3mm*
- *Weight = 65kg*

Which treatment is best **for me**?

- Which anti-hypertensive treatment?
  - Calcium channel blocker (A)
  - ACE inhibitor (B)

- Future blood pressure: treatment A vs. B
- **Individualized Treatment Effect (ITE)**

Which treatment is best **for me**?

- Which anti-hypertensive treatment?
  - Calcium channel blocker (A)
  - ACE inhibitor (B)

- Potential *confounder*: maybe rich patients got medication A more often, and poor patients got medication B more often

Example 2
Job training:
Average Treatment Effect (ATE)

Should the government fund job-training programs?

- Existing job training programs seem to help unemployed and underemployed find better jobs
- Should the government fund such programs?
- Maybe training helps but only marginally? Is it worth the investment?
- **Average Treatment Effect (ATE)**
- Potential *confounder*: Maybe only motivated people go to job training? Maybe they would have found better jobs anyway?

Observational studies

A major challenge in causal inference from observational studies is how to *control* or *adjust for* the confounding factors

Counterfactuals and causal inference

- Does treatment T cause outcome Y?
- If T had not occurred, Y would not have occurred (David Hume)

- Counterfactuals:
  Kim received job training (T), and her income one year later (Y) is 20,000$
  What would have been Kim's income had she not had job training?

Counterfactuals and causal inference

- Counterfactuals:
  Kim received job training (T), and her income one year later (Y) is $20,000
  What would have been Kim's income had she not had job training?
- If her income would have been $18,000, we say that job training caused an increase of $2,000 in Kim's income
- The problem: you never know what might have been

Sliding Doors

Potential Outcomes Framework (Rubin-Neyman Causal Model)

- Each unit 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***"
- Individual Treatment Effect for unit i:

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

- Average Treatment Effect:

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

Potential Outcomes Framework (Rubin-Neyman Causal Model)

- Each unit 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:

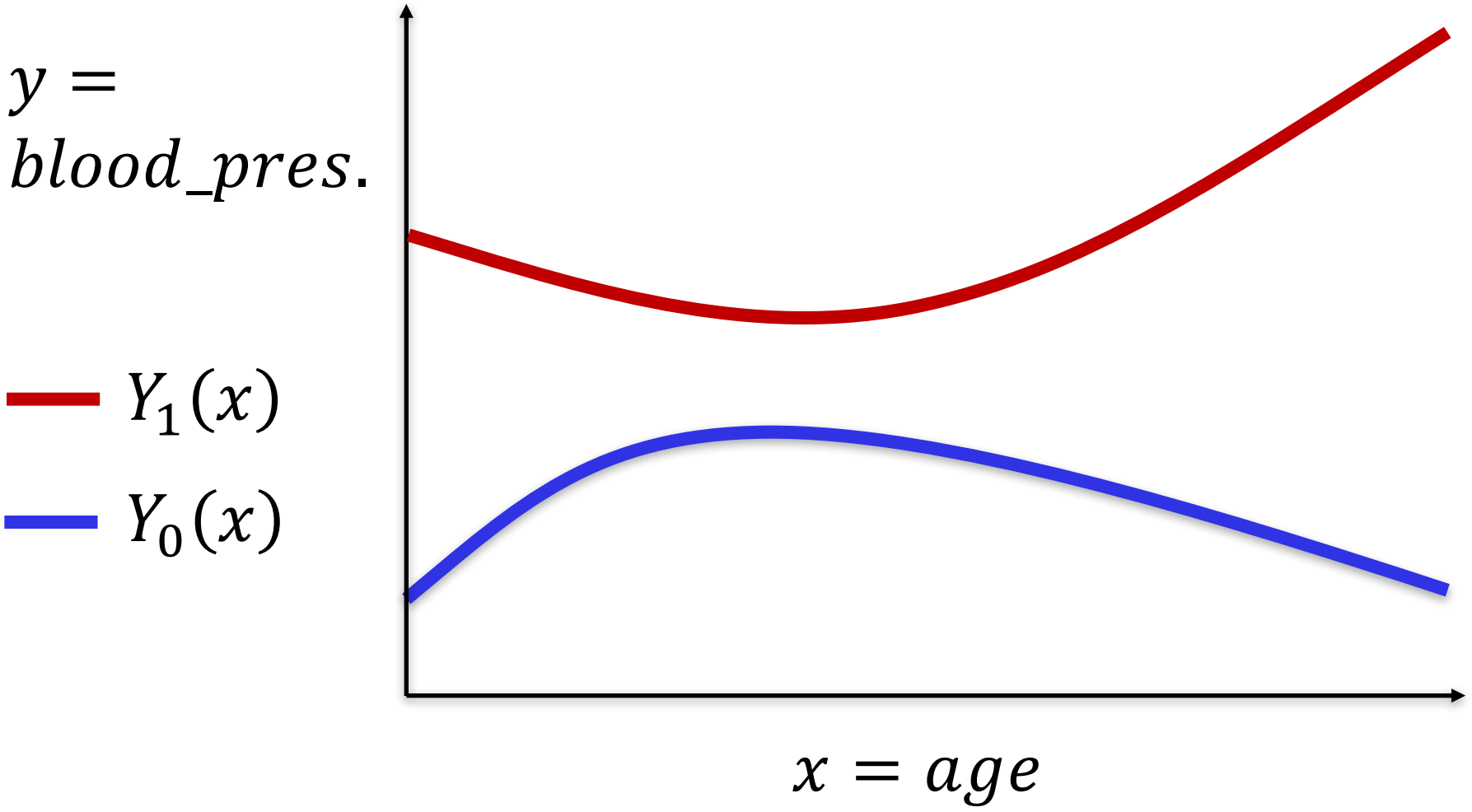
$$y_i^{CF} = (1 - t_i) Y_1(x_i) + t_i Y_0(x_i)$$

Terminology

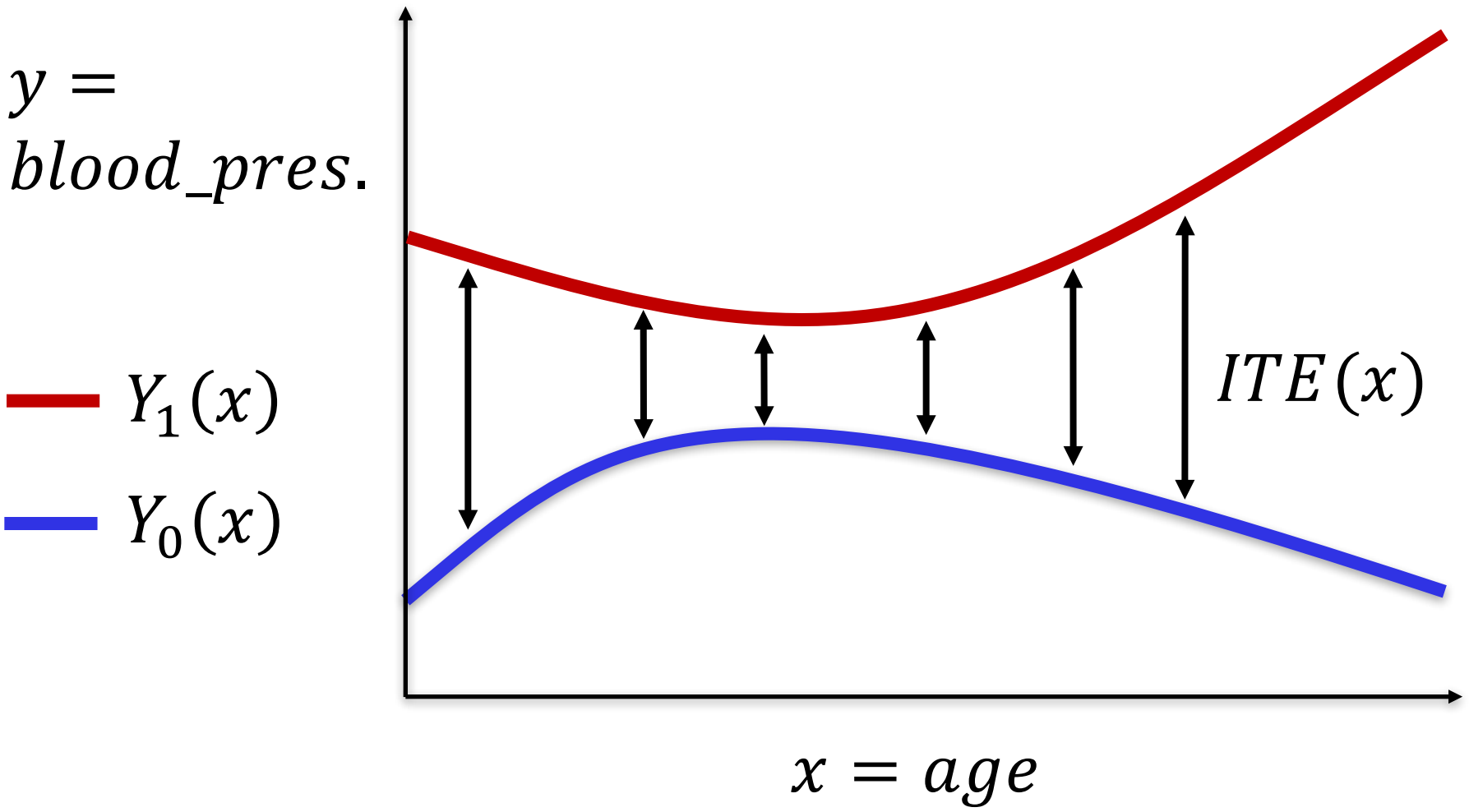
- *Unit*: data point, e.g. patient, customer, student
- *Treatment*: binary indicator (in this tutorial)
  Also called *intervention*
- *Treated*: units who received treatment=1
- *Control*: units who received treatment=0
- *Factual*: the set of observed units with their respective treatment assignment
- *Counterfactual*: the factual set with flipped treatment assignment

Example – Blood pressure and age

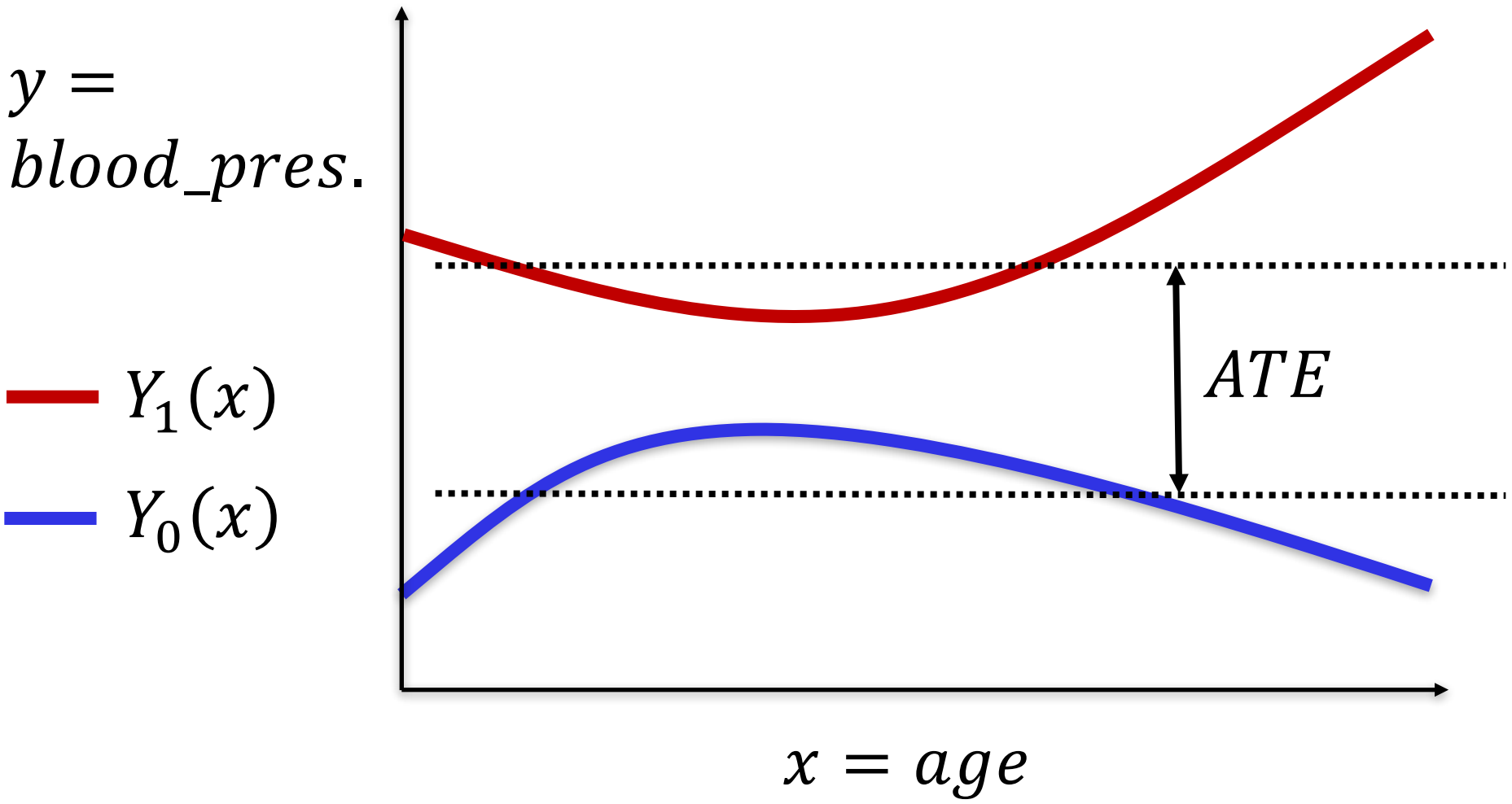
$y = blood\_pres.$



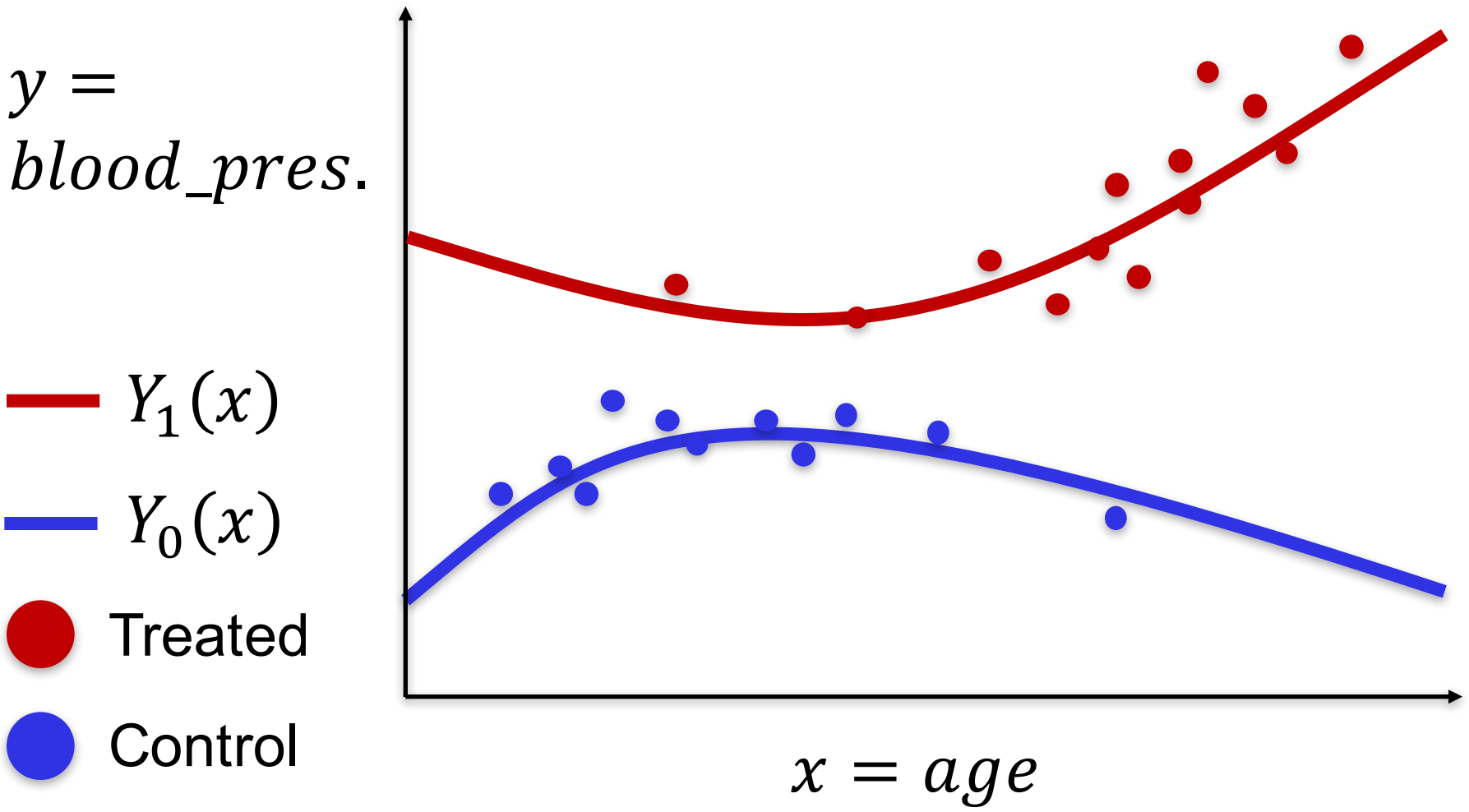
Blood pressure and age



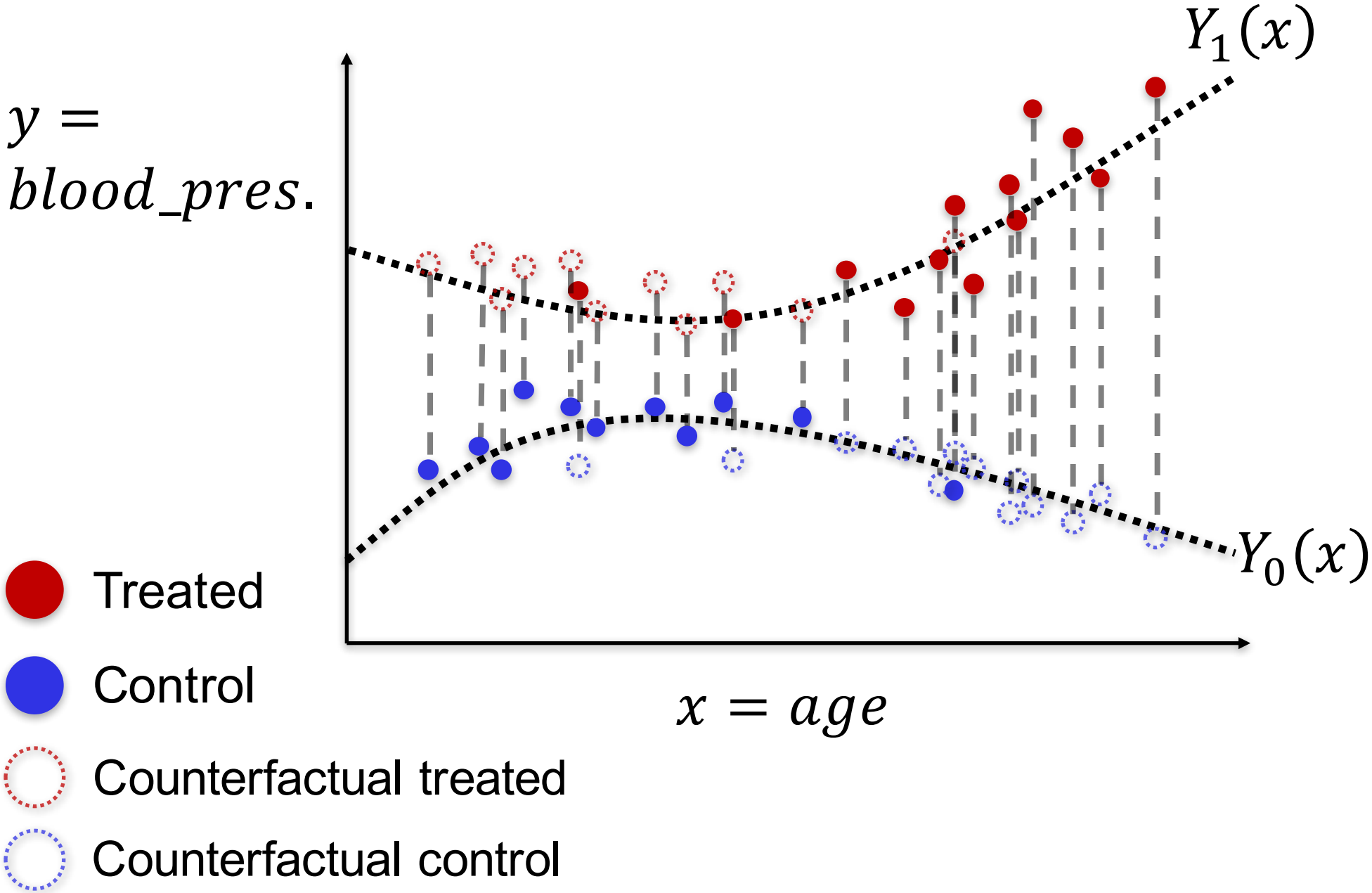
Blood pressure and age



Blood pressure and age



Blood pressure and age



“The fundamental problem of causal inference”

We only ever observe one of the two outcomes

"The Assumptions" – 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

"The Assumptions" – no unmeasured confounders

Y_0, Y_1: potential outcomes for control and treated

x: unit covariates (features)

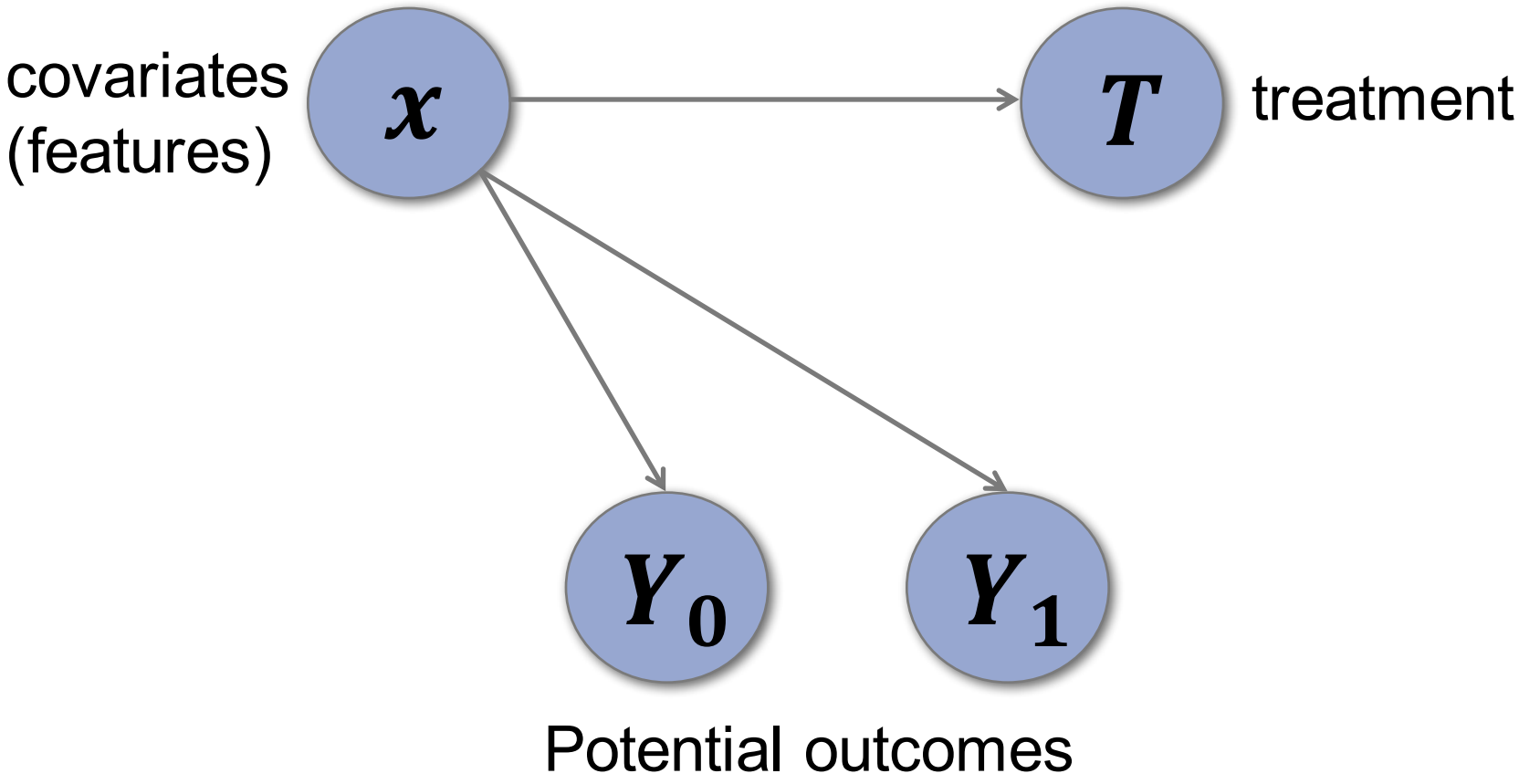
T: treatment assignment

We assume:

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

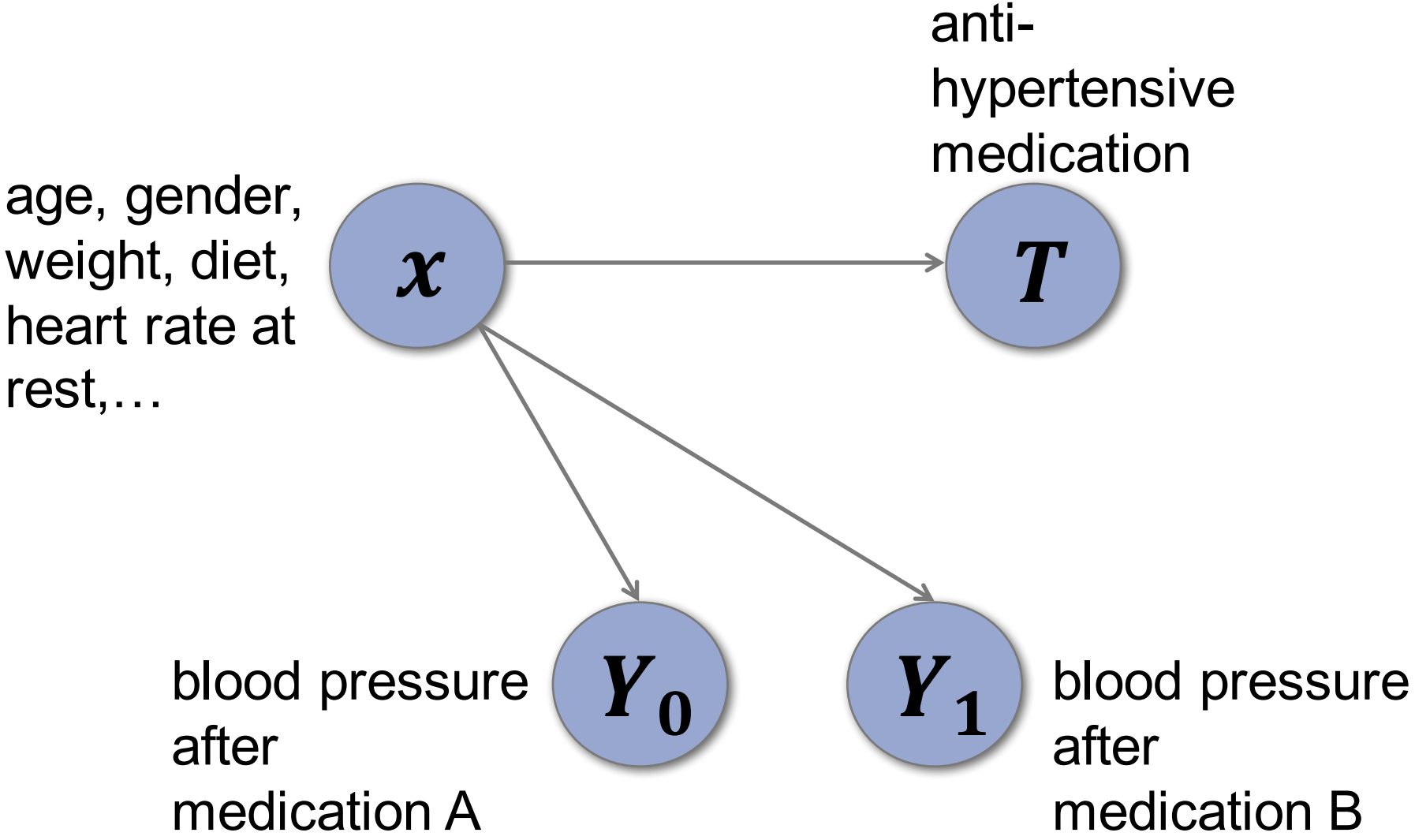
Ignorability

Ignorability



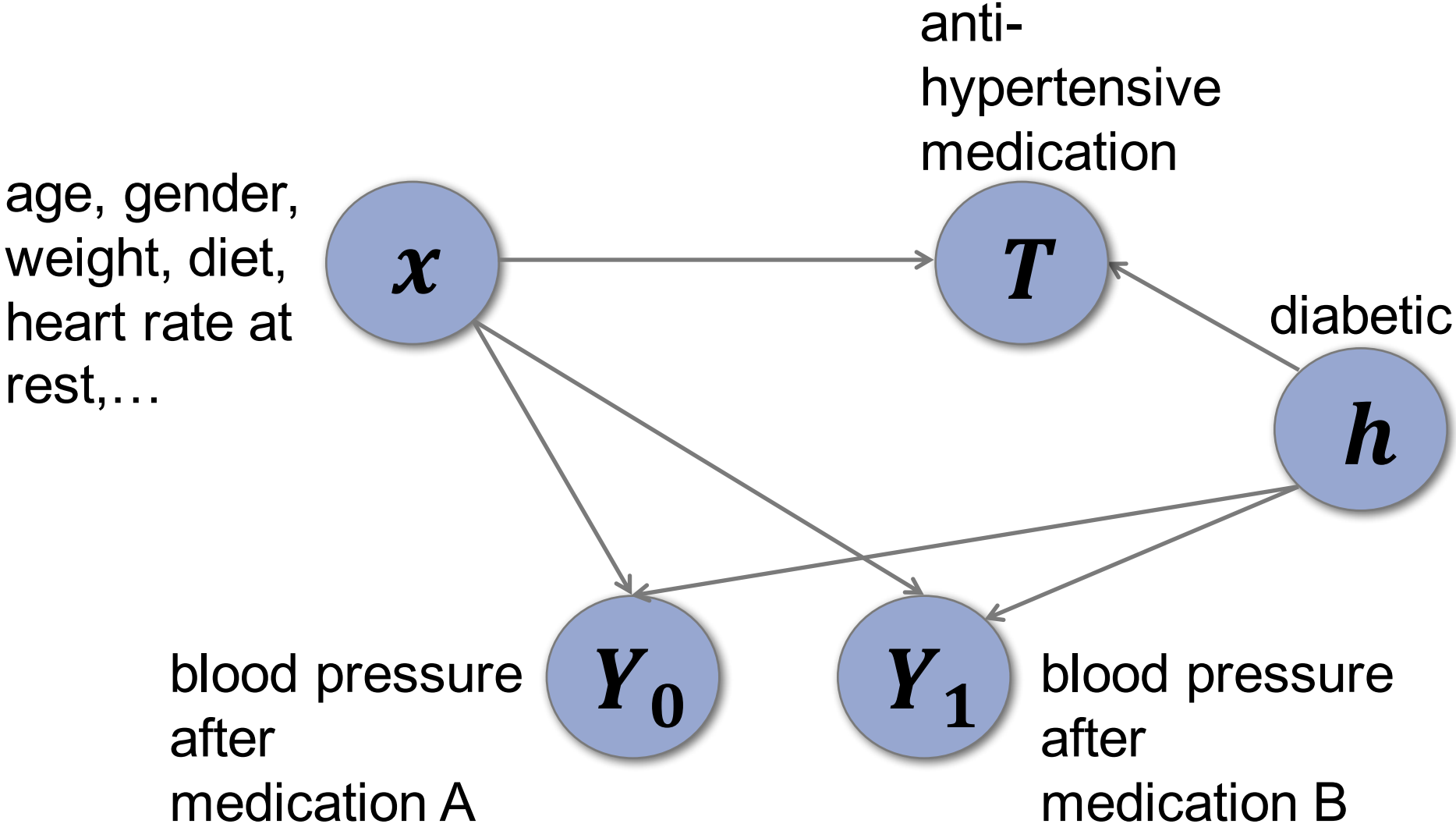
$$(Y_0, Y_1) \perp\!\!\!\perp T \mid x$$

Ignorability



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

No Ignorability



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

“The Assumptions” – common support

Y_0, Y_1: potential outcomes for control and treated

x: unit covariates (features)

T: treatment assignment

We assume:

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

Average Treatment Effect

The expected causal effect of T on Y:

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

Average Treatment Effect – the adjustment formula

- Assuming ignorability, we will derive the *adjustment formula* (Hernán & Robins 2010, Pearl 2009)
- The adjustment formula is extremely useful in causal inference
- Also called *G-formula*

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:

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

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

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

ignorability

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

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

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:

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

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

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

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

} 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:

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

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

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

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

} Quantities we can estimate from data

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

Outline

Tools of the trade

- Matching
- Covariate adjustment
- Propensity score

Set up

- Samples: $x_1, x_2, \ldots, x_n$
- Observed binary treatment assignments: $t_1, t_2, \ldots, t_n$
- Observed outcomes: $y_1, y_2, \ldots, y_n$

$$x = (age, gender, married, education, income\_last\_year, \ldots)$$
$$t = (no\_job\_training, job\_training)$$
$$y = income\_one\_year\_after\_training$$

- Does job training raise average future income?

Outline

Tools of the trade

- Matching
- Covariate adjustment
- Propensity score

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

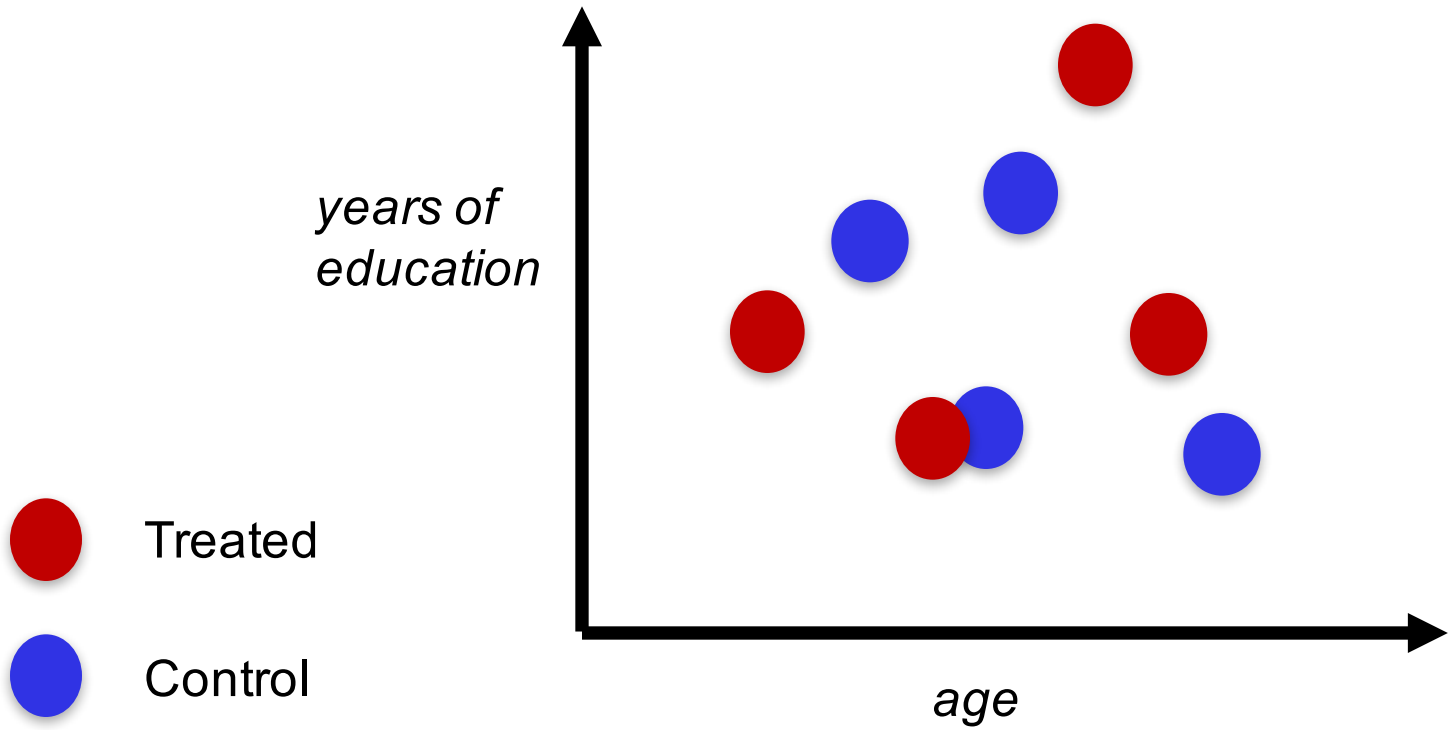
Obama, had he gone to law school

Obama, had he gone to business school

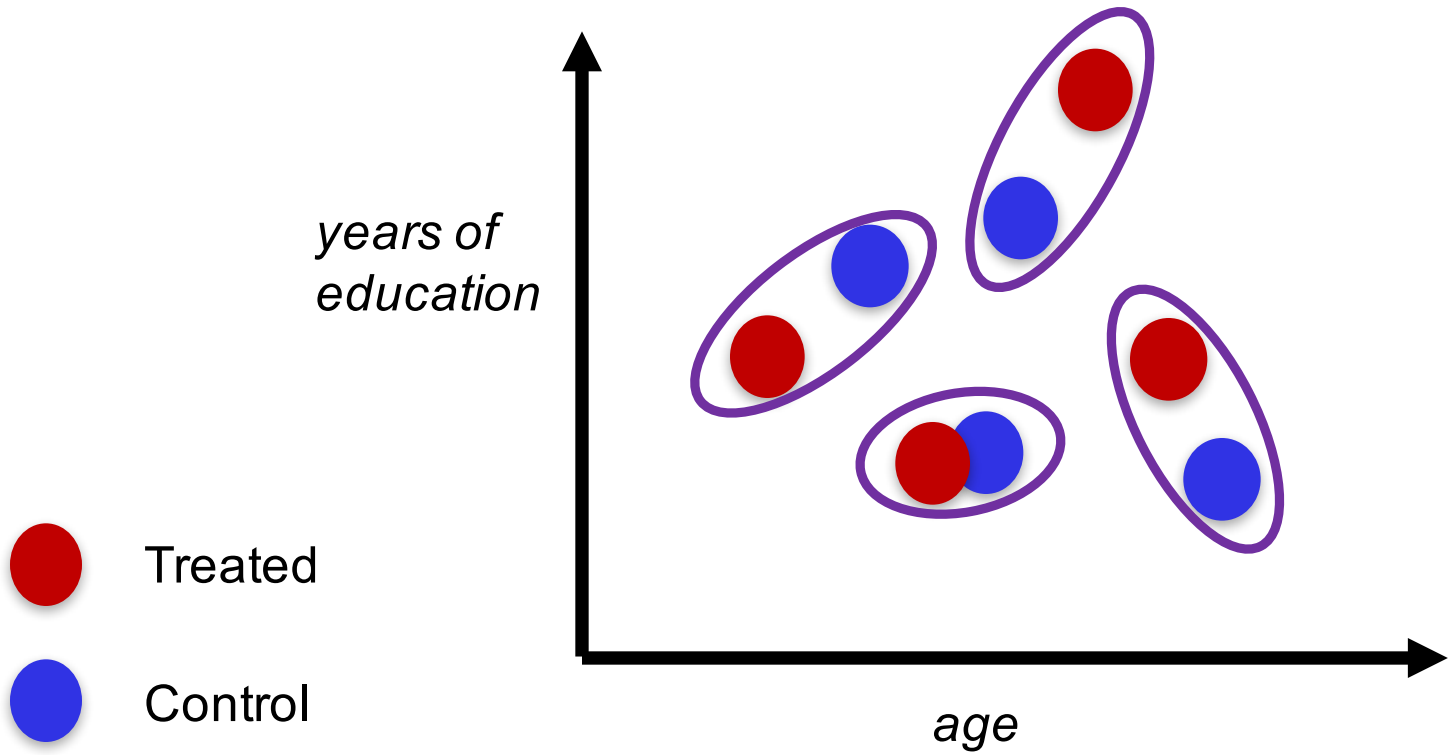
Matching

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

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:

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

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

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

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

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

Matching

- Interpretable, especially in small-sample regime
- Nonparametric
- Heavily reliant on the underlying metric (however see below about propensity score matching)
- Could be misled by features which don't affect the outcome

Matching

- Many other matching methods we won't discuss:
  - Coarsened exact matching
    Iacus et al. (2011)
  - Optimal matching
    Rosenbaum (1989,2002)
  - Propensity score matching
    Rosenbaum & Rubin (1983), Austin (2011)
  - Mahalanobis distance matching
    Rosenbaum (1989,2002)

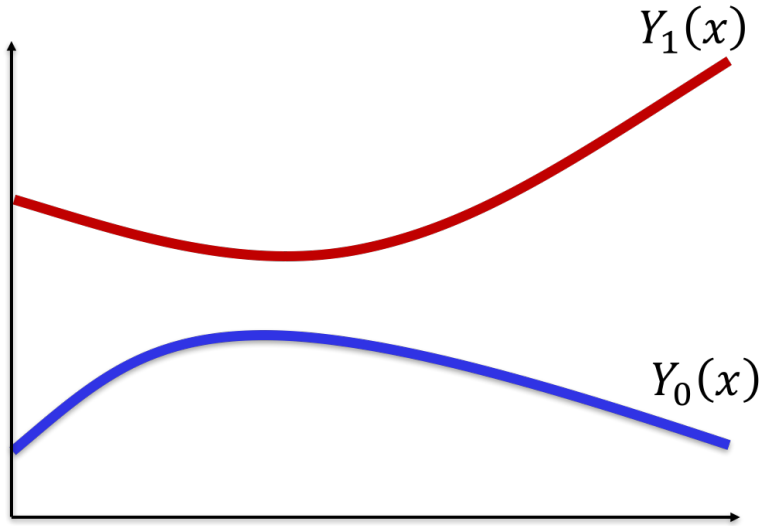
Outline

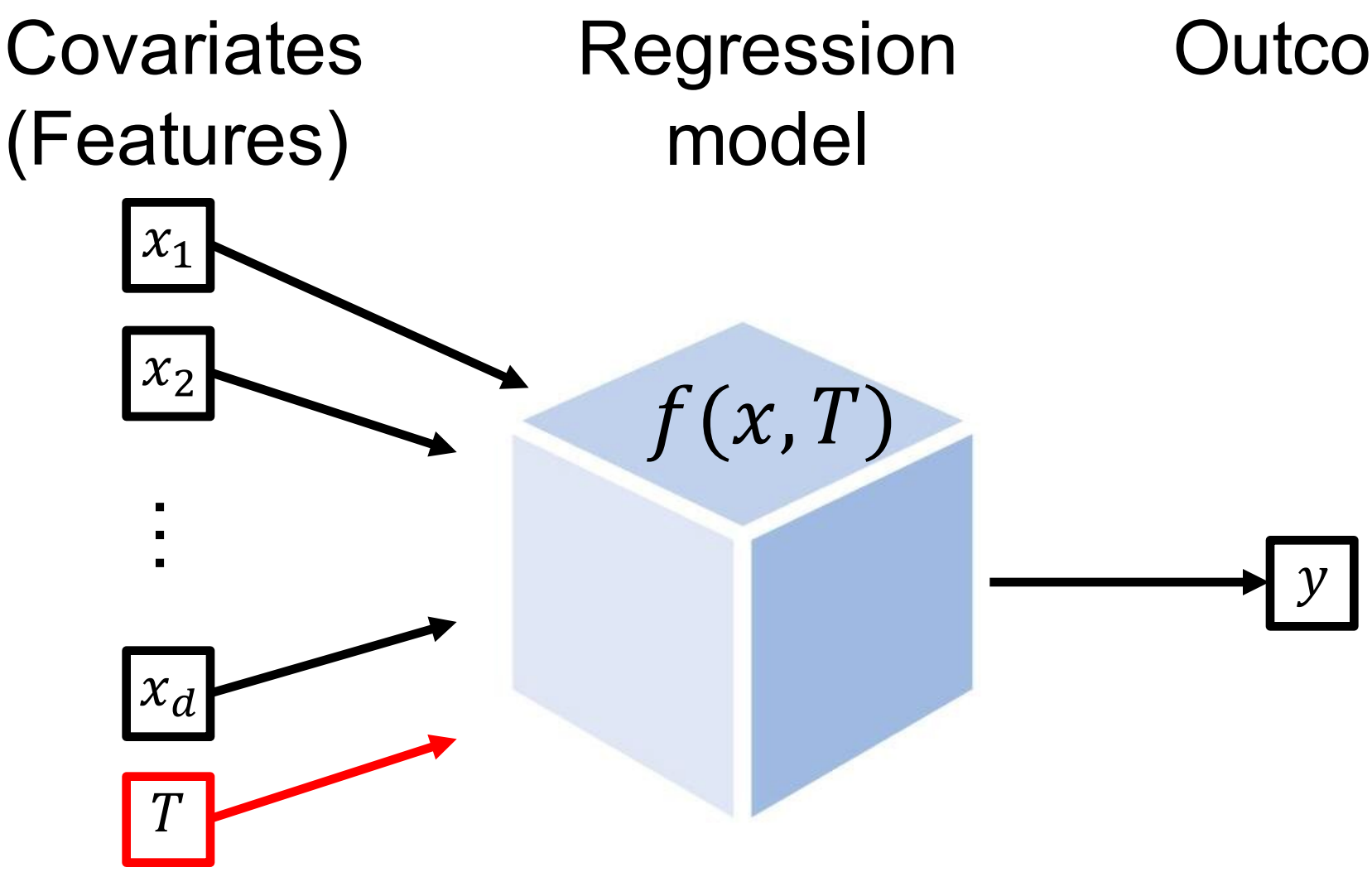
Tools of the trade

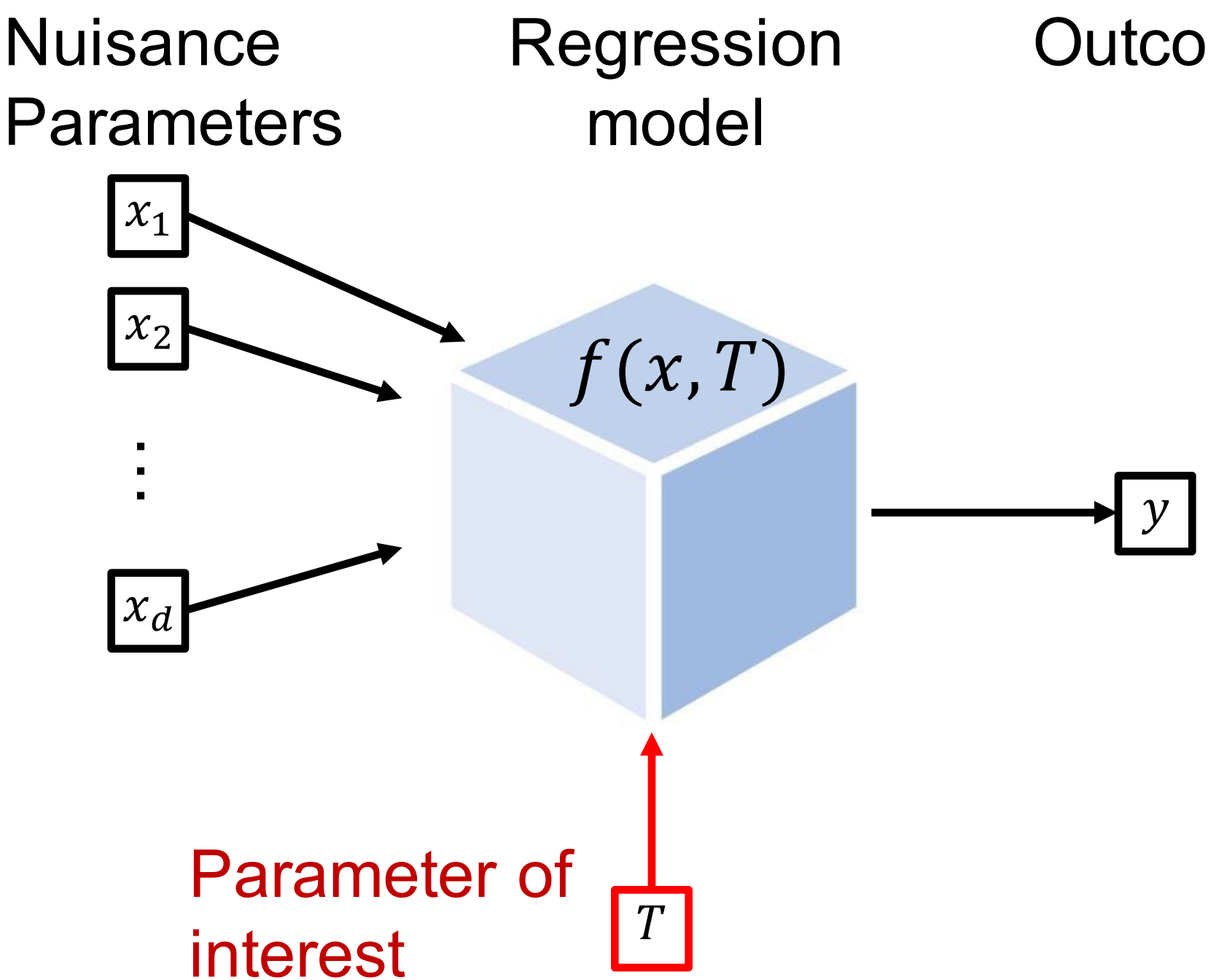
- Matching
- Covariate adjustment
- Propensity score

Covariate adjustment

- Explicitly model the relationship between treatment, confounders, and outcome
- Also called “Response Surface Modeling”
- Used for both ITE and ATE
- A regression problem




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
T
Parameter of interest

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)

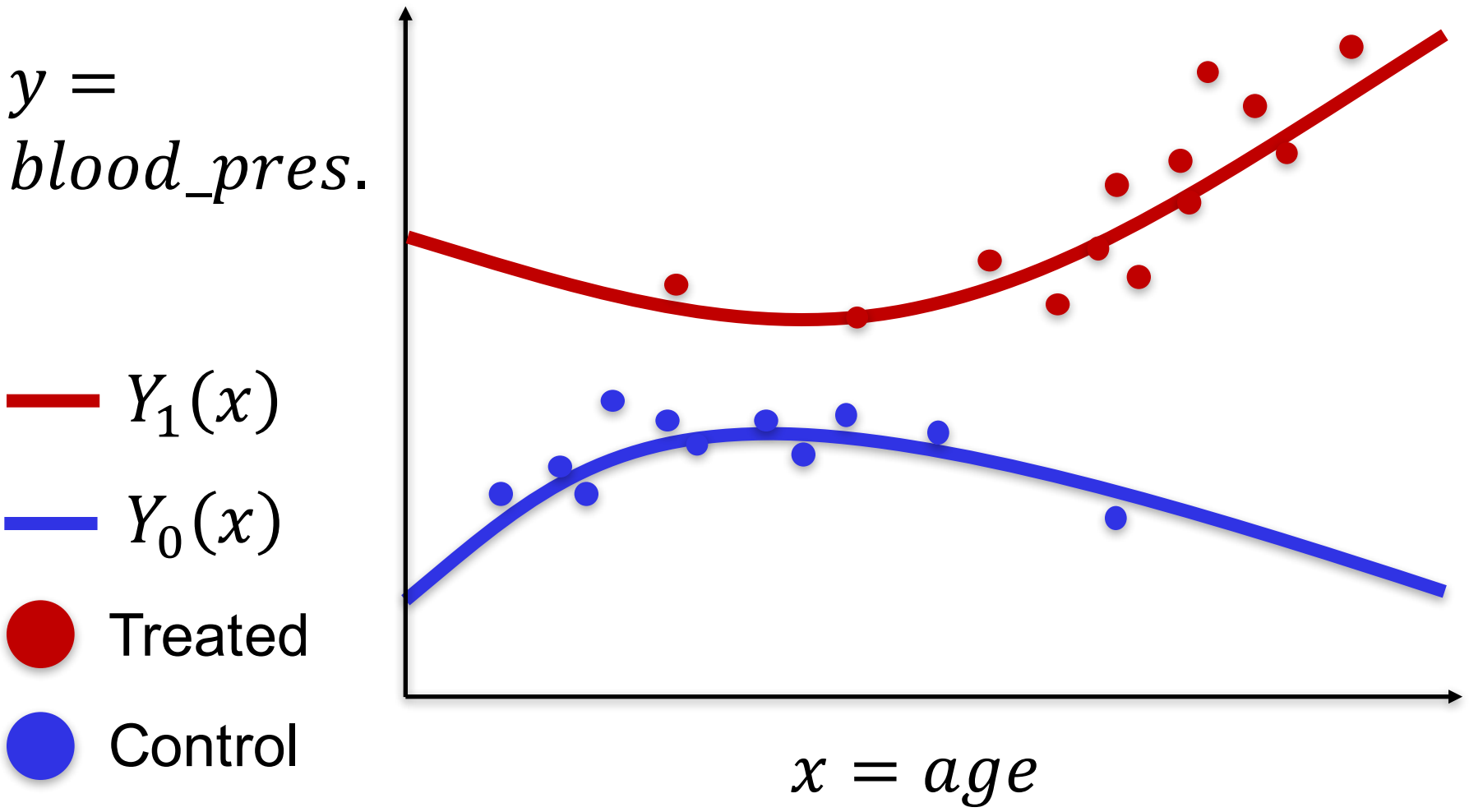
- Explicitly model the relationship between treatment, confounders, and outcome
- Under ignorability, the expected causal effect of T on Y:

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

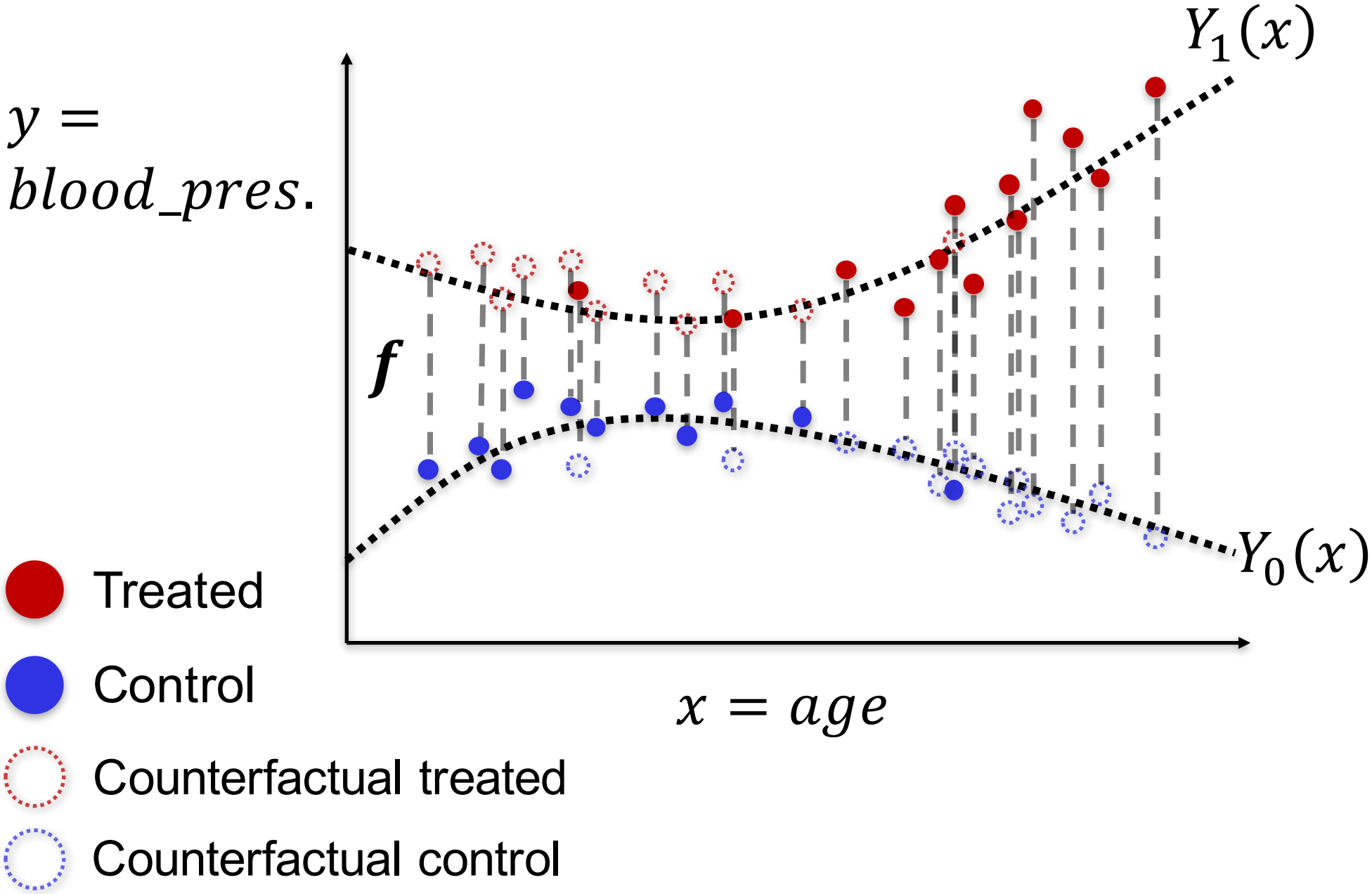
- Fit a model $f(x,t)\approx\mathbb{E}[Y_t|T=t,x]$

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

Covariate adjustment



Covariate adjustment



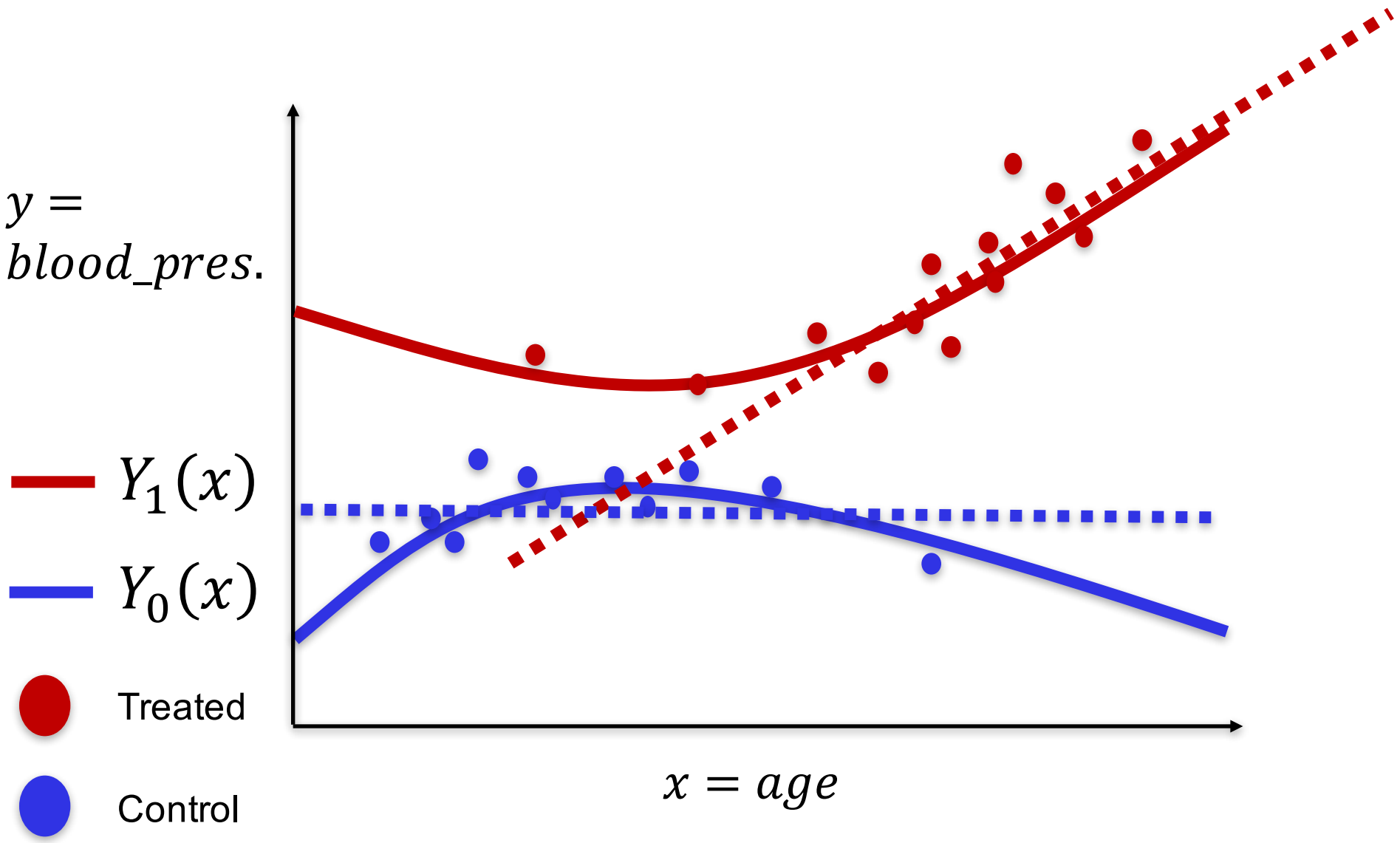
Warning: this is not a classic supervised learning problem

- Our model was optimized to predict outcome, not to differentiate the influence of A vs. B
- What if our high-dimensional model threw away the feature of medication A/B?
- Maybe the model never saw a patient like Anna get medication A? Maybe there's a reason patients like Anna never get A?

Covariate adjustment - consistency

- If the model $f(x,t) \approx \mathbb{E}[Y_t|T=t,x]$ is consistent in the limit of infinite samples, then under ignorability the estimated $\widehat{ATE}$ will converge to the true ATE
- A sufficient condition: overlap and well-specified model

Covariate adjustment: no overlap



Linear model

- Assume that:

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

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

- Then:

$$ITE(x) := Y_1(x) - Y_0(x) =$$

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

Linear model

- Assume that:

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

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

- We care about γ, not about $Y_t(x)$
  Identification, not prediction

Linear model

blood pressure **age,weight,…** **medication**

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

  Hypertension is affected by many variables: lifestyle, weight, genetics, age
- Each of these often stronger **predictor** of blood-pressure, compared with type of medication taken
- Regularization (e.g. Lasso) might remove the treatment variable!
- Features → (“nuisance parameters”, “variable of interest”)

Regression - 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:

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

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

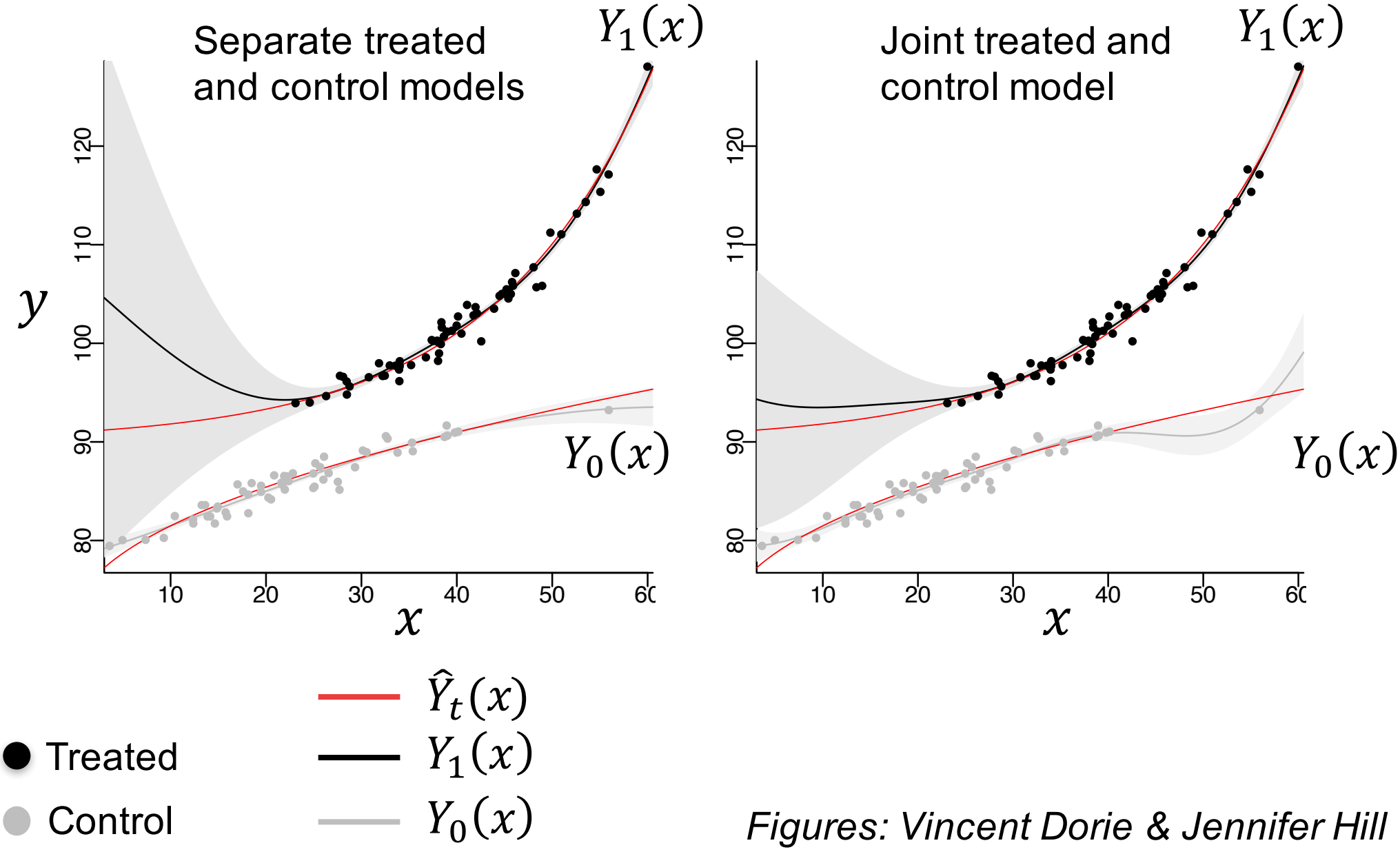
Using machine learning for causal inference

- Machine learning techniques can be very useful and have recently seen wider adoption
- Random forests and Bayesian trees
  Hill (2011), Athey & Imbens (2015), Wager & Athey (2015)
- Gaussian processes
  Hoyer et al. (2009), Zigler et al. (2012)
- Neural nets
  Beck et al. (2000), Johansson et al. (2016), Shalit et al. (2016), Lopez-Paz et al. (2016)
- "Causal" Lasso
  Belloni et al. (2013), Farrell (2015), Athey et al. (2016)

Using machine learning for causal inference

- Machine learning techniques can be very useful and have recently seen wider adoption
- How is the treatment variable used:
  - Fit two different models for treated and control?
  - Not regularized?
  - Privileged

Example: Gaussian process



Figures: Vincent Dorie & Jennifer Hill

Covariate adjustment and matching

- Matching is equivalent to covariate adjustment with two 1-NN 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)

Outline

Tools of the trade

- Matching
- Covariate adjustment
- Propensity score

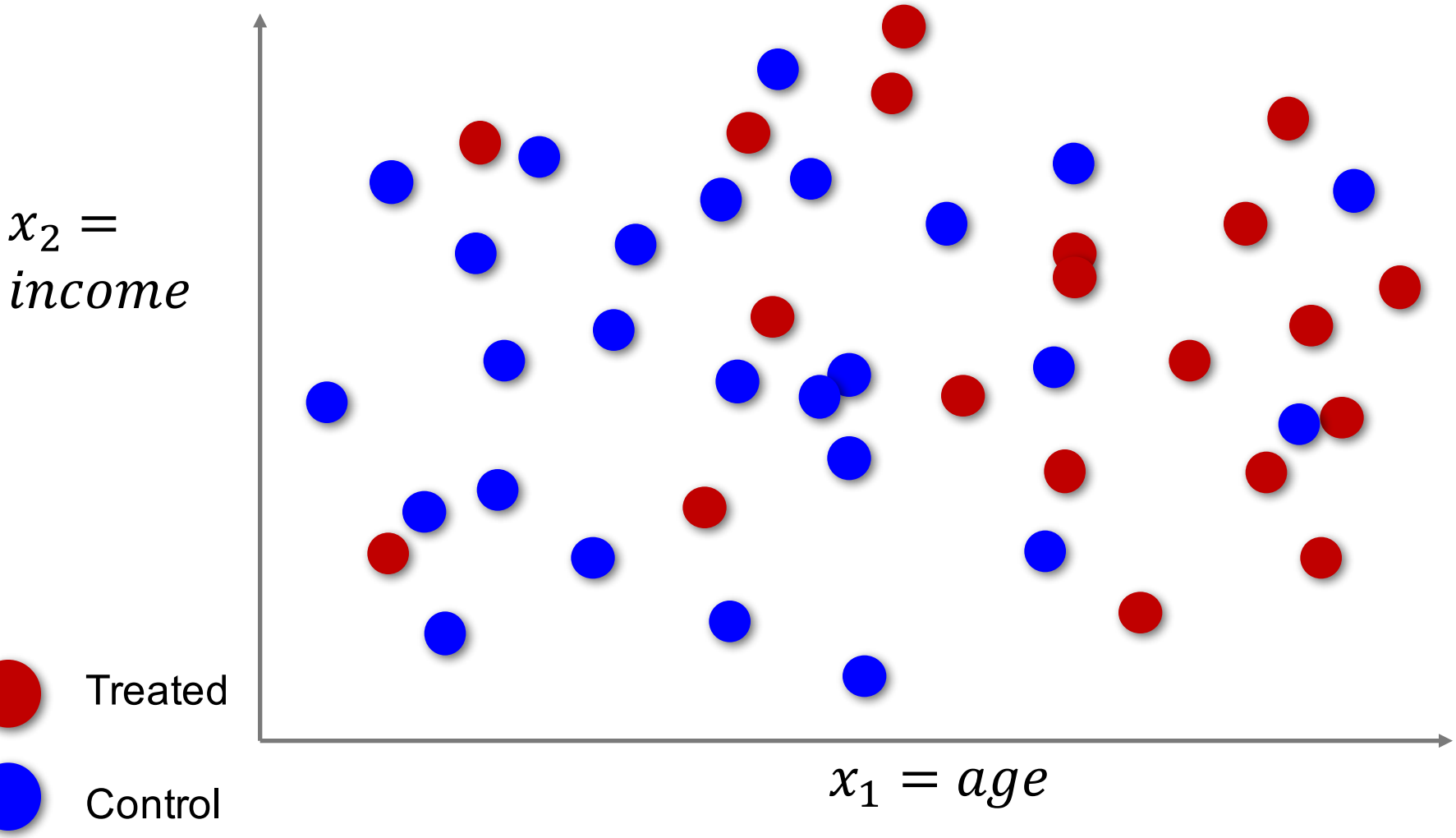
Propensity score

- 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

How to obtain ATE with propensity score

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 $p = 0.5$

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

- Recall average treatment effect:

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

- We only have samples for:

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

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

Propensity scores - derivation

- We only have samples for:

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

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 only have samples for:

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

- We want:

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

- We know that:

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

- Then:

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

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

Calculating the propensity score

- If $p(T = t|x)$ is known, then propensity scores re-weighting is consistent
- Example: ad-placement algorithm samples $T = t$ based on a known algorithm
- Usually the score is unknown and must be estimated
- Example: use logistic regression to estimate the probability that patient x received medication $T = t$
- Calibration: must estimate the probability correctly, not just the binary assignment variable

“The Assumptions” – ignorability

- If ignorability doesn’t hold then the average treatment effect is **not** $\mathbb{E}_{x\sim p(x)}\left[\mathbb{E}[Y_1|T=1,x]-\mathbb{E}[Y_0|T=0,x]\right]$, invalidating the starting point of the derivation

“The Assumptions” – overlap

- If there’s not much overlap, propensity scores become non-informative and easily miscalibrated

- Sample variance of inverse propensity score re-weighting scales with $\sum_{i=1}^{n} \frac{1}{\hat{p}(T=1|x_i)\hat{p}(T=0|x_i)}$, which can grow very large when samples are non-overlapping
(Williamson et al., 2014)

Propensity score in machine learning

- Same idea is in importance sampling!
- Used in off-policy evaluation and learning from logged bandit feedback (Swaminathan & Joachims, 2015)
- Similar ideas used in covariate shift work (Bickel et al., 2009)