Causal Analysis: A Quick Intro

Part 2: Introduction to Machine Learning for Causal Analysis using Observational Data

CAUSAL ANALYSIS USING DATA FROM OBSERVATIONAL STUDIES

- We want to estimate the *causal effect* of treatment or social exposure T on outcome Y
  - Causal effect is policy-relevant: what benefits accrue if we intervene to change T?
  - Treatment must be *modifiable* for this to make sense – otherwise, what's the point??

- We have data from an **observational** study where T and Y are measured
  - How were the individual units in the data set collected?
  - Which population were these units drawn from?
  - Temporal ordering: are we sure treatment was determined before outcome? **If not, game over!**

REGRESSION ESTIMATION

- Linear regression is workhorse for effect estimation
  - For subject i, we observe their treatment t_i and outcome y_i and fit the model

$$y_i = a + bt_i + \epsilon_i$$

where we focus on binary treatment

$$t_i = \begin{cases} 1 & \text{if } i \text{ received treatment} \\ 0 & \text{control} \end{cases}$$

  - Coefficient b is the **difference between the mean outcomes in the treatment and control groups**
  - Usually estimate using ordinary least squares or maximum likelihood

Note: Can elaborate regression model if treatment is continuous

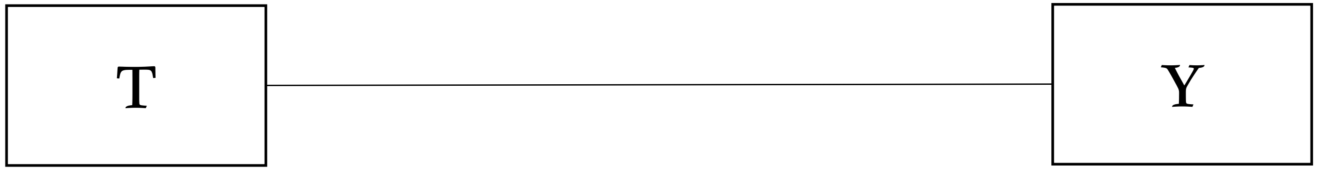
e.g. Add t_i^2, t_i^3, t_i^4 , etc. terms (curvilinear) or use dummy variables (stepwise linear) to capture more complex relationships in a limited way

CONFOUNDING IN OBSERVATIONAL STUDIES

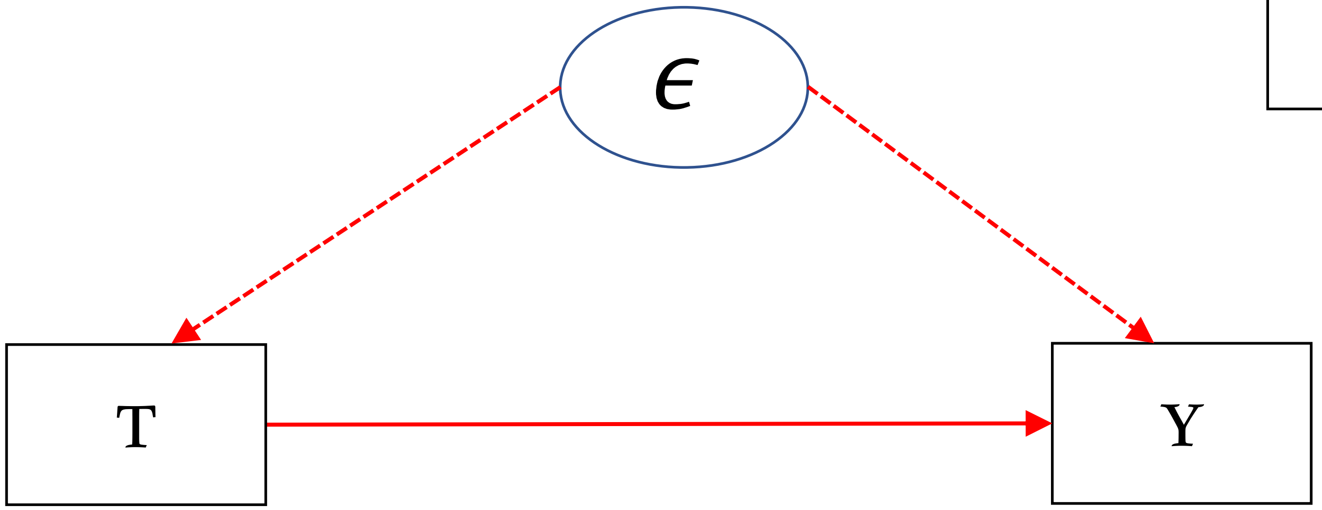
- Regression coefficient b is a measure of association between T and Y
  - Would equal *causal effect* if RCT data (randomised controlled trial) where T was randomized

- But T not randomised: treatment selected **in a way that could depend (indirectly**) on Y
  - Same 'type' of person who chooses treatment is also the sort who has high outcome (& vice versa)
  - Banks give loans to people more likely to successfully pay off their loans
  - Children from wealthier families more likely to attend private school and have better post-school outcomes
- Would have done better anyway: association *confounds* this with the true effect of treatment

1. Graph for association



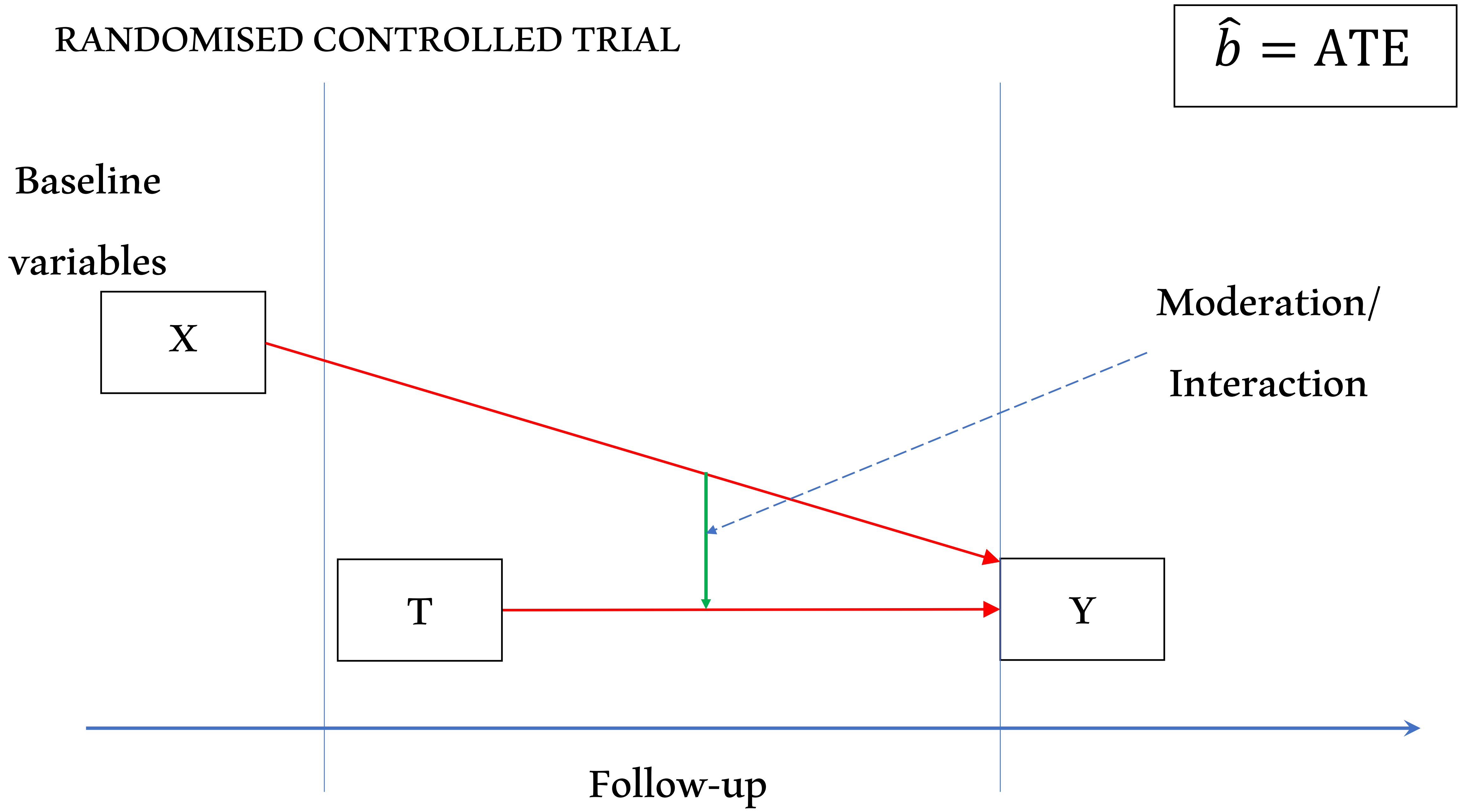
2. Causal graph



$$\boxed{\hat{b} \neq \text{ATE}}$$

ROLE OF BASELINE VARIABLES

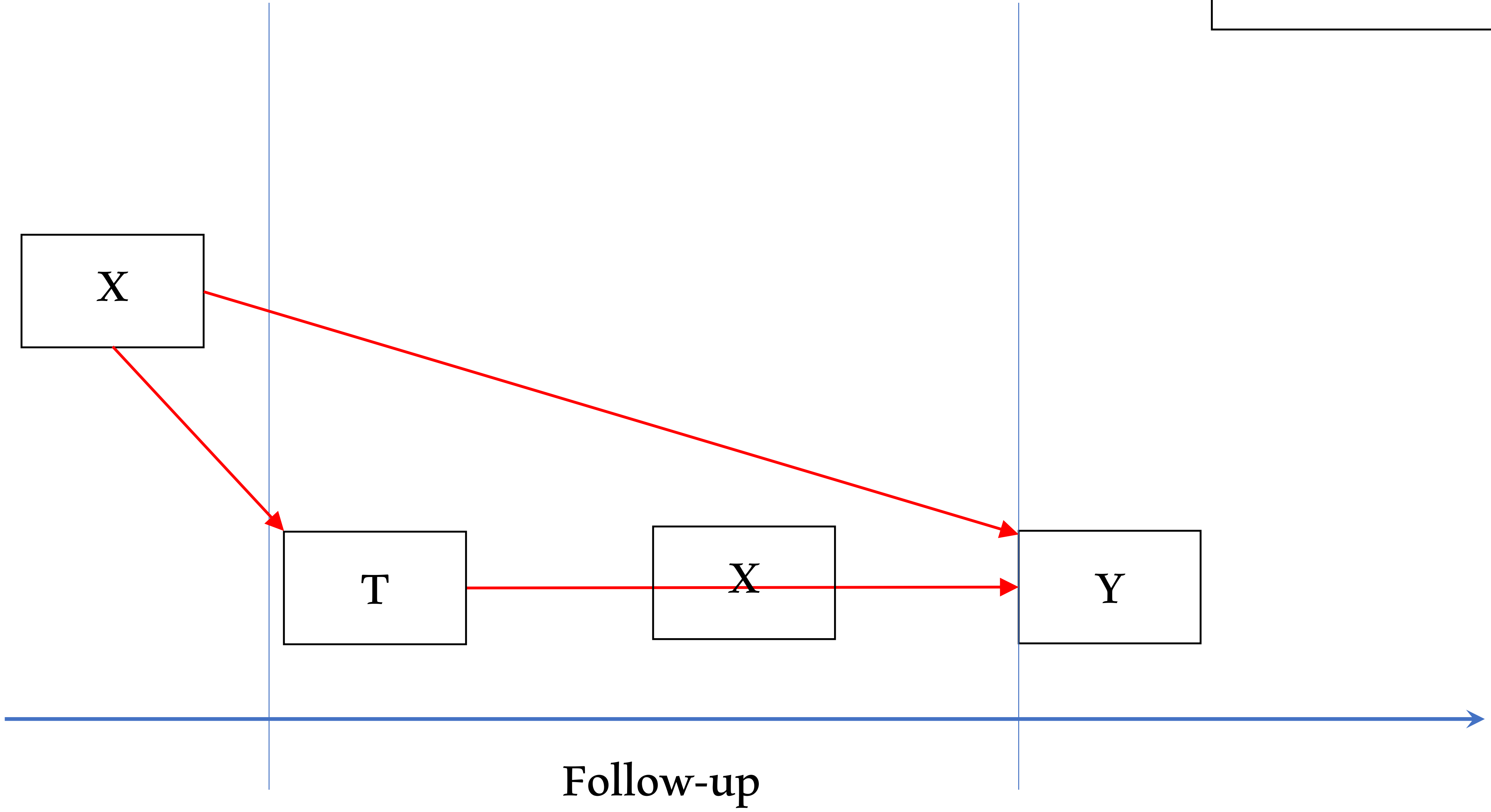
- Suppose study measures many other variables $X = (X_1, X_2, \ldots, X_p)$
- Throw away those we know happened after the treatment was chosen
  - Not a baseline variable if so!
  - We need to be sure we have a quasi-experimental study
- Distribution of X generally different for treated and untreated in observational study



RANDOMISED CONTROLLED TRIAL
$\hat{b} = \text{ATE}$
Baseline variables
X
Moderation/ Interaction
T
Y
Follow-up

OBSERVATIONAL STUDIES AS QUASI-EXPERIMENTS

$$\boxed{\hat{b} \neq \text{ATE}}$$



CAUSAL EFFECT: THE AVERAGE TREATMENT EFFECT (ATE)

- Potential outcomes
  - For subject i, we observe their treatment t_i and outcome y_i and fit the model

$$\mathcal{Y}_i^0, \mathcal{Y}_i^1$$

where we observe only one

$$y_i = \begin{cases} \mathcal{Y}_i^1 & \text{if } i \text{ received treatment} \\ \mathcal{Y}_i^0 & \text{control} \end{cases}$$

- The average/mean of $\mathcal{Y}_i^1 - \mathcal{Y}_i^0$ across everyone in the *target population*

$$\text{ATE} = \frac{1}{N}\sum_i (\mathcal{Y}_i^1 - \mathcal{Y}_i^0) = E[\mathcal{Y}_i^1 - \mathcal{Y}_i^0]$$

Notes: If treatment T is polytomous or continuous then $\mathcal{Y}_i^t$ is a set of values so need model for effect of treatment as many different ways of measuring treatment effect

Implicitly assume stable unit treatment value assumption (SUTVA): potential outcomes don't depend on what other units get

IGNORABLE SELECTION

Independent/Uncorrelated

- Treatment selection is (strongly) ignorable if

$$\begin{pmatrix} \mathcal{Y}_i^1 \\ \mathcal{Y}_i^0 \end{pmatrix} \perp\!\!\!\perp t_i \Big| x_i$$

  - Differences between treated and untreated among those subjects characterized by same X are **random**

- Referred to as **no unobserved confounding** or **no omitted variables**

- The challenges now are

  - Verifying this is true [clue: *We can't! But must do what we can to mitigate confounding*]

  - Adjusting the estimate of b to account for these effects [Today's focus!]

Other approaches needed if there is unobserved confounding (e.g. instrumental variables) but beyond scope

Weakly ignorable $\mathcal{Y}_i^0 \perp\!\!\!\perp t_i | x_i$ --- generally estimate ATE *among the treated*: ATT $= E[\mathcal{Y}_i^1 - \mathcal{Y}_i^0 | t_i = 1]$

CONDITIONAL AVERAGE TREATMENT EFFECT

- Introducing X also introduces the conditional average treatment effect (CATE)

$$\text{CATE}(x_i) = E\left[\mathcal{Y}_i^1 - \mathcal{Y}_i^0 \middle| x_i\right]$$

- ATE is simply the average value of $\text{CATE}(x_i)$ in the population:

$$\text{ATE} = E\{\text{CATE}(x_i)\}$$

- Also written as difference between conditional means:

$$\text{CATE}(x_i) = E\left[\mathcal{Y}_i^1 \middle| x_i\right] - E\left[\mathcal{Y}_i^0 \middle| x_i\right] = \mu_1(x_i) - \mu_0(x_i)$$

  - We exploit this later on…

REGRESSION ADJUSTMENT

- Include X variables in the regression model

$$y_i = a + bt_i + cx_i + \epsilon_i$$

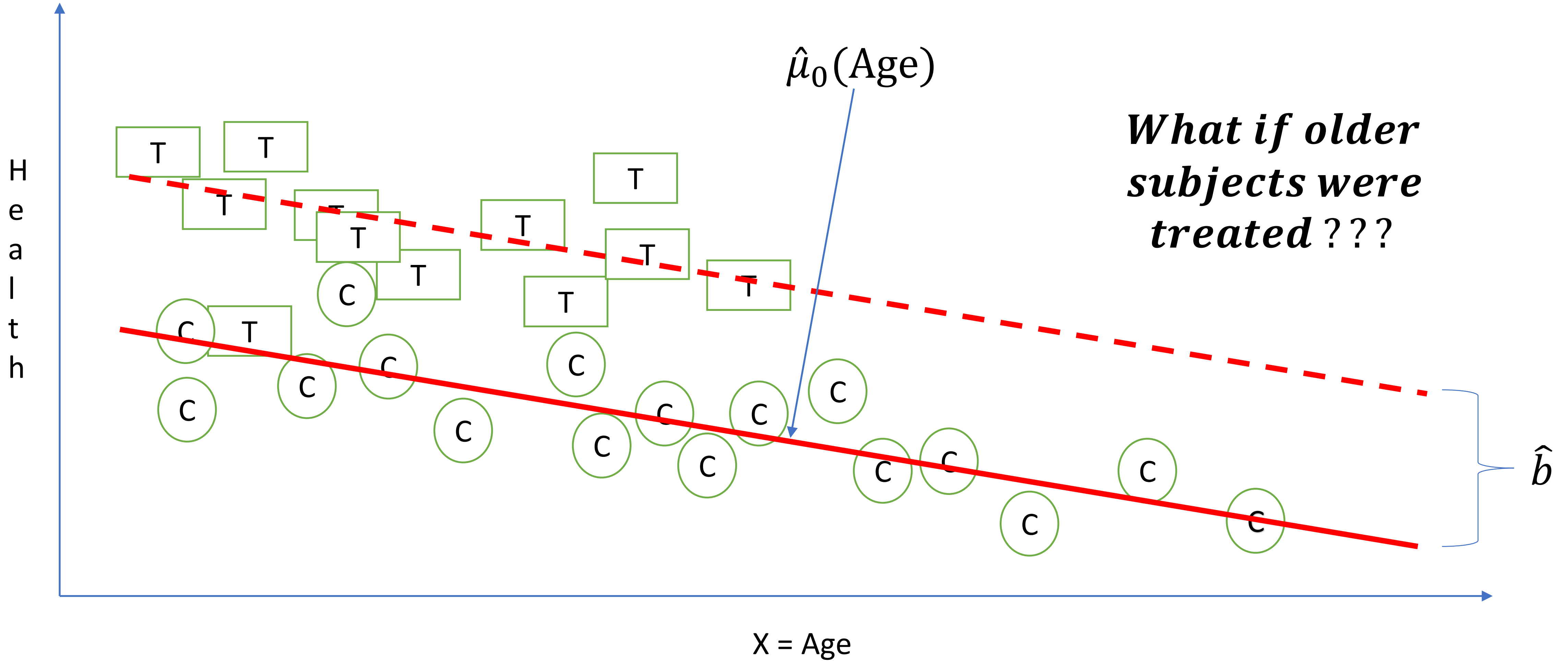
where $cx_i = c_1 x_{1i} + \cdots + c_p x_{pi}$ is linear combination of the confounding variables

- Regression works if the mean of untreated potential outcomes is linear, that is,

$$\mu_0(x_i) = E\left[\mathcal{Y}_i^0 \middle| x_i\right] = a + cx_i$$

- <u>... and</u> the causal effects are **homogeneous**
  - This means that $\mathrm{CATE}(x_i)$ is a constant that does not depend on x_i
  - Simply add b to get the treated mean: $\mu_1(x_i) = \mu_0(x_i) + b$
- 'Extrapolates' whenever there is *no overlap* in the data (see next slide)

COVARIATE (NO) OVERLAP



INVERSE PROBABILITY WEIGHTING

> Specify same model for treatment (it excludes X) as if data were from RCT

$$y_i = a + bt_i + \varepsilon_i$$

> Handle selection by estimating *selection propensities*

$$\Pr[t_i = 1 | x_i] = e(x_i)$$

> > No longer assume homogeneous effects

> > $e(x) = 0$ or 1 implies **no overlap** - makes clear we cannot estimate $\mathrm{CATE}(x)$

Estimate using logistic regression

> Balance the sample by fitting a weighted regression with weights

$$w_i = \frac{t_i}{\hat{e}(x_i)} + \frac{1 - t_i}{1 - \hat{e}(x_i)} = \begin{cases} 1/\hat{e}(x_i) & \text{for treated} \\ 1/1 - \hat{e}(x_i) & \text{for untreated} \end{cases}$$

Note. Ignorable assumption needed to ensure that $\Pr\left[t_i = 1 | \mathcal{Y}_i^0, \mathcal{Y}_i^1, x_i\right] = \Pr[t_i = 1 | x_i]$

Selection propensities can also play a key role for matching estimators (to estimate 'counterfactual' $\mathcal{Y}_i^{1-t_i}$ to match 'factual' y_i)

(SUPERVISED) MACHINE LEARNING (RECAP)

- Algorithms that learn the true relationship between *input variables* and *output variables*
  - Set up to accurately predict outputs/outcomes
  - We call different ML algorithms *base learners* or just *learners*
  - Earlier discussed regression classification, decision trees, random forests
- Differences
  - Move away from parametric models $\mathrm{Y} = f(X; \theta)$, just learn rule $f: X \rightarrow Y$
  - Move from statistical model selection to *train, validate* and *test* (incl. setting *meta-parameters*)
  - Results in predicted outcomes rather than parameter estimates

Note. Even regression classification, linear model simply device for prediction, do not need to believe it is true

META-ALGORITHMS FOR ESTIMATING ATE: S-, T- AND X-LEARNERS

> Use the power of ML to estimate causal effects more accurately

> Learn $\mu_t(x_i)$ or $e(x_i)$ or both (depends on which estimator you choose)

> Different learning strategies for ATE: S – single, T – two, X – hybrid strategy

> Then plug in learnt $\mu_t(x_i)$ or $e(x_i)$ to a valid estimators of ATE:

$$\widehat{\text{ATE}}_{\text{PredDiff}} = \frac{1}{n}\sum_i \{\hat{\mu}_1(x_i) - \hat{\mu}_0(x_i)\}$$

$$\widehat{\text{ATE}}_{\text{IPW}} = \frac{1}{n}\sum_i \left\{\frac{t_i}{\hat{e}(x_i)} - \frac{1-t_i}{1-\hat{e}(x_i)}\right\} y_i = \frac{1}{n}\sum_i w_i y_i$$

$$\widehat{\text{ATE}}_{\text{Pred-IPW}} = \frac{1}{n}\sum_i w_i \hat{\mu}_{t_i}(x_i)$$

DOUBLY ROBUST ESTIMATION

> Theoretically,* the most accurate (low bias, small variance) estimator of all

> Combines strengths of $\hat{\mu}_t(x_i)$ or $\hat{e}(x_i)$ (although $\widehat{\text{ATE}}_{\text{Pred-IPW}}$ does this too)

> Also allows for biased estimation of either $\hat{\mu}_t(x_i)$ or $\hat{e}(x_i)$ but not both (no free lunch!)

> Simply combine three estimators from previous slides:

$$\widehat{\text{ATE}}_{\text{DR}} = \widehat{\text{ATE}}_{\text{PredDiff}} + \widehat{\text{ATE}}_{\text{IPW}} - \widehat{\text{ATE}}_{\text{Pred-IPW}}$$

---

* Theoretically - simply means in large samples; its performance in small-to-medium sized samples is less clear-cut

S-LEARNERS

We'll use random forest regression

- Learn single structural model from **all available data**

$$s(t,x) = E[y_i | t_i = t, x_i = x] + error$$

- Then estimate

$$\hat{\mu}_0(x_i) = s(0, x_i) \text{ and } \hat{\mu}_1(x_i) = s(1, x_i)$$

- Compared with linear regression:
  - Allows $\mu_0(x_i) = E\left[\mathcal{Y}_i^0 \middle| x_i\right]$ to be non-linear
  - and heterogenous treatment effects

T-LEARNERS

Random forest regression

- T: Two stages, one for treated, one for untreated:
  - From treated units, learn $s_1(x_i) = E[y_i|t_i = 1, x_i] + error$
  - From control units, learn $s_0(x_i) = E[y_i|t_i = 0, x_i] + error$

- Then

$$\hat{\mu}_0(x_i) = s_0(x_i) \text{ and } \hat{\mu}_1(x_i) = s_1(x_i)$$

META-ALGORITHMS: X-LEARNER

> Combines learner predictions with observed data:

> As with T-learner, learn $s_1(x_i)$ from treated units and $s_0(x_i)$ from untreated units

> Also learn selection propensity $e(x)$ using suitable base learner

> Calculate 'imputed' individual treatment effects

$$D_i = \begin{cases} D_i^1 \coloneqq y_i - s_0(x_i) & \text{if } i \text{ is treated} \\ D_i^0 \coloneqq s_1(x_i) - y_i & \text{if } i \text{ is untreated} \end{cases}$$

> Apply base learner to untreated and treated groups:

a) Learn $\hat{\tau}_0(x) = E\left[D_i^0 \middle| x_i = x\right]$ and $\hat{\tau}_1(x) = E\left[D_i^1 \middle| x_i = x\right]$
b) Calculate $\widehat{\mathrm{CATE}}(x) = \hat{\tau}_0(x)\big(1 - \hat{e}(x)\big) + \hat{\tau}_1(x)\hat{e}(x)$

> Estimate of ATE is simply sample average of $\widehat{\mathrm{CATE}}(x)$

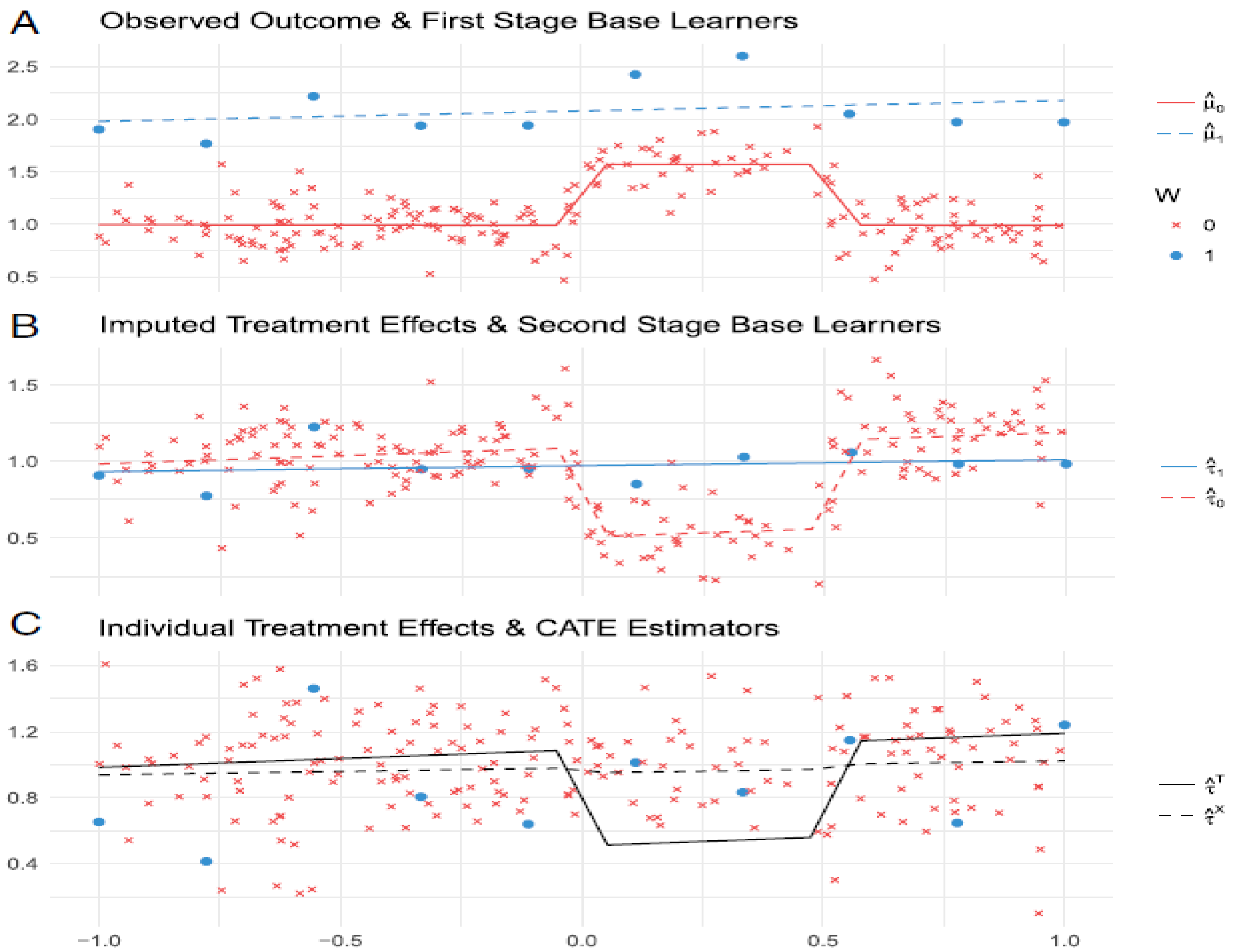


Fig. 1. Intuition behind the X-learner with an unbalanced design. (*A*) Observed outcome and first-stage base learners. (*B*) Imputed treatment effects and second-stage base learners. (*C*) ITEs and CATE estimators.

From Kuntzel at al. (2019) Metalearners for estimating heterogeneous treatment effects using machine learning.

SOME REFERENCES AND FURTHER READING

- Wager and Athey (2018) [identifying heterogenous treatment effects with random forests]

  https://doi.org/10.1080/01621459.2017.1319839

- Econ-ML repository [research papers on ML in economics – there's a lot of work going on!]

  http://econ-neural.net/

- Hernan and Robins (2020) [exhaustive book on causal analysis]

  https://cdn1.sph.harvard.edu/wp-content/uploads/sites/1268/2020/02/ci_hernanrobins_21feb20.pdf

- Kuntzel et al (2019) [X-learners vs S- and T-learners]

  https://doi.org/10.1073/pnas.1804597116

- Xu et al (2020) [where the computer scientists are headed…]

  https://arxiv.org/abs/2006.16789

Now to the practical bit

But first, any questions…?