Causal Analysis: A Quick Intro

Day 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

$$y_i = a + bt_i + e_i$$

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

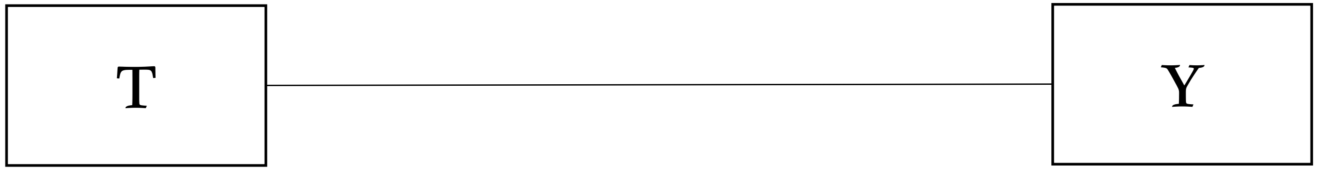
e.g. Add t_i^2, t_i^3, t_i^4 , etc. terms (curvilinear) or use dummy variables (stepwise linear) to capture more complex relationships in a limited way

CONFOUNDING IN OBSERVATIONAL STUDIES

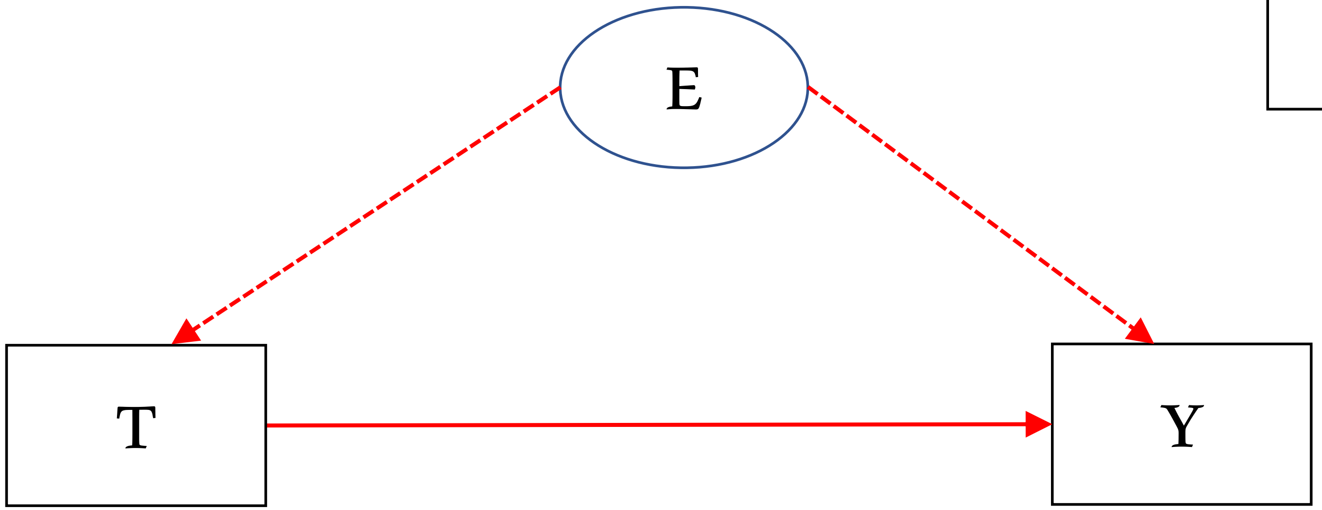
- Regression coefficient b is a measure of association between T and Y
  - Would equal *causal effect* if RCT data (randomised controlled trial) where T was randomized

- 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



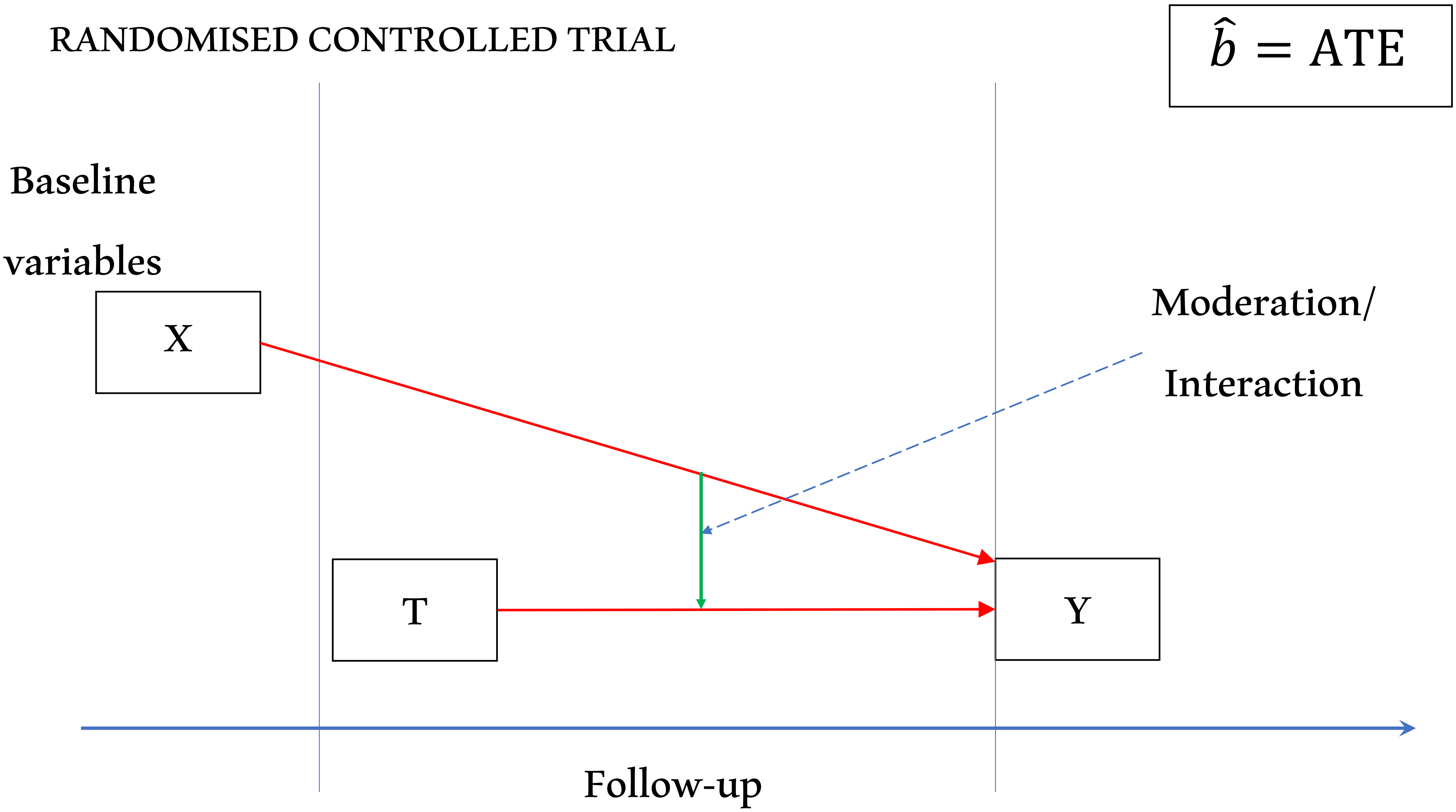
$$\boxed{\hat{b} \neq \text{ATE}}$$

ROLE OF BASELINE VARIABLES

- Suppose study measures many other variables $X = (X_1, X_2, \dots, X_p)$

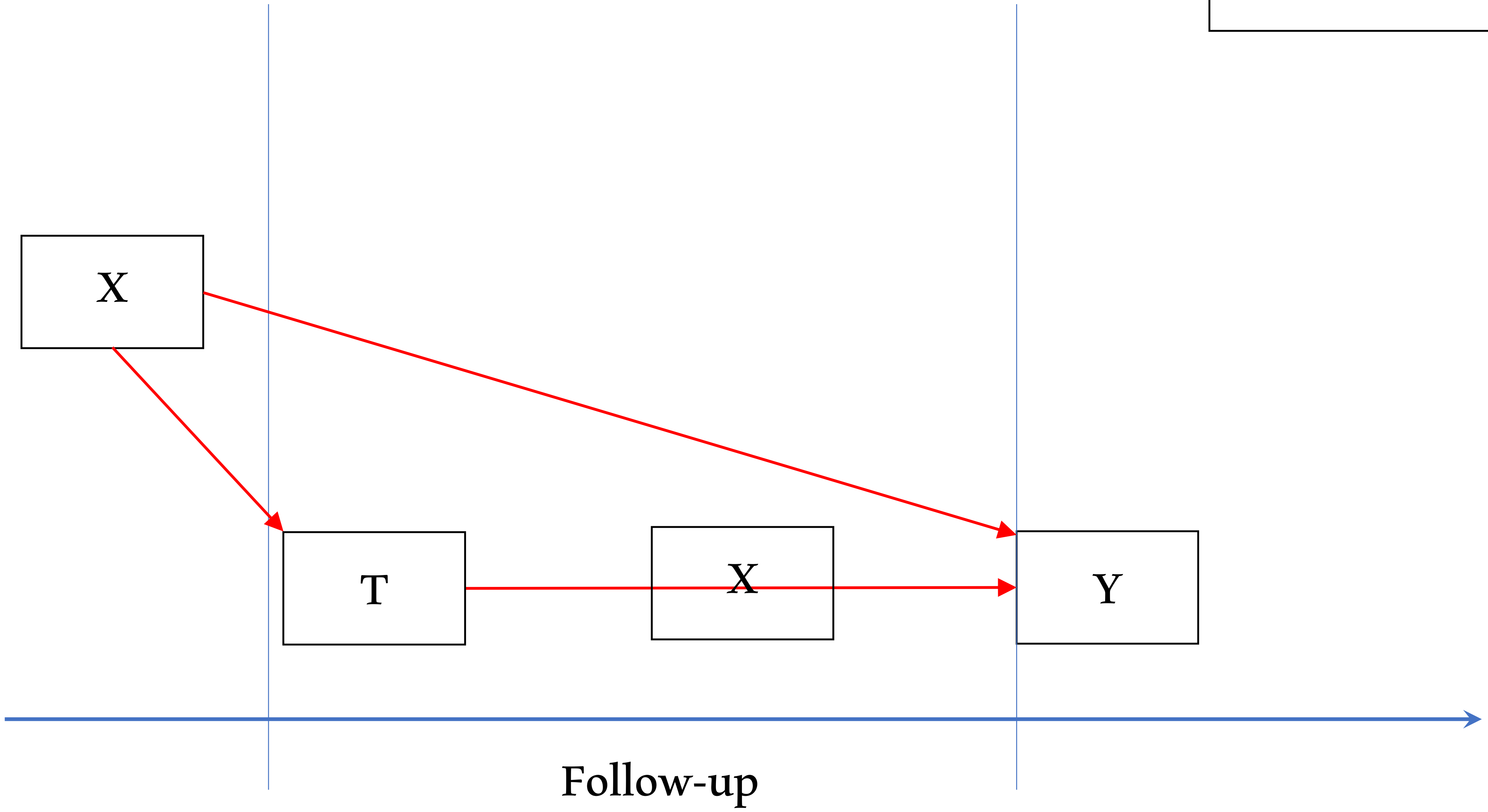
- Throw away those we know happened after the treatment was chosen
  - Not a baseline variable if so!
  - We need to be sure we have a quasi-experimental study

- Distribution of X generally different for treated and untreated in observational study


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

OBSERVATIONAL STUDIES AS QUASI-EXPERIMENTS

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



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*

$$\text{ATE} = E\left[\mathcal{Y}_i^1 - \mathcal{Y}_i^0\right]$$

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** assumption
- The challenges now are
  - Verifying this condition is true [clue: you can't! Doing what you 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 $= E\left[\mathcal{Y}_i^1 - \mathcal{Y}_i^0 | t_i = 1\right]$

REGRESSION ADJUSTMENT

- Include X variables in the regression model

$$y_i = a + bt_i + cx_i + e_i$$

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

- This models mean of the untreated potential outcomes as linear model

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

  - Fit driven entirely by data on untreated subjects
  - Assumes causal effects are **homogeneous**
  - Usually performs well with small number of X variables (especially if categorical)
  - Extrapolates if no overlap are allowed (but predicted under the model)

Homogeneous effects if **Conditional** ATE $\mathrm{CATE}(x_i) = E\left[\mathcal{Y}_i^1 - \mathcal{Y}_i^0 \middle| x_i\right] = \mathrm{ATE}$

No overlap for e.g. $x =$ (young, male, low SEP, unhealthy) either if all units are treated or all units are untreated

INVERSE PROBABILITY WEIGHTING

> Specify (marginal) structural model for treatment (it excludes X)

$$y_i = a + bt_i + e_i$$

> But these have to go somewhere – into the *selection propensities*

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

> No longer assume homogeneous effects

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

> IPW estimator - weighted regression using (weighted sample is 'balanced')

$$w_i = \frac{t_i}{\hat{e}(x_i)} + \frac{1 - t_i}{1 - \hat{e}(x_i)}$$

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

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)

IPW estimator is not fully efficient but Robins's **doubly robust** estimator is, and also robust to mis-specification of either propensity or structural model (but not both)

(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*
  - Yesterday looked at regression classification, decision trees, random forests
- Differences
  - Move away from parametric models $\mathrm{Y} = f(X;\theta)$, just $f: X \to Y$
  - Move from statistical model selection to *train* 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 learners from ML to estimate causal effects more accurately

> Target the Conditional Average Treatment Effect (CATE)

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

where

$$\mu_t(x_i) = E\left[\mathcal{Y}_i^t \middle| x_i\right]$$

> Different estimation strategies: S – single, T – two-estimator strategy, X – hybrid strategy

> Then take average, e.g.

$$\text{ATE} = \frac{1}{N}\sum_{i=1}^{N} \text{CATE}(x_i)$$

S-LEARNERS

We'll use random forest regression

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

$$\mu(t, x) = E[y_i | t_i = t, x_i = x]$$

> Then estimate

$$\text{CATE}(x_i) = \mu(1, x_i) - \mu(0, x_i),$$

> Compared with linear regression:

> Not limited to linear model for $\mu_0(x_i) = E\left[\mathcal{Y}_i^0 | x_i\right]$

> Allows heterogenous treatment effects

ANOTHER S-ESTIMATOR: ‘DOUBLY ROBUST’ IPW ESTIMATION

Random forest classifier

> Learn propensity model $\Pr[t_i = 1 | x_i = x] = e(x)$
  > Simply fit **weighted** regression of Y on T using $w_i = t_i/\hat{e}(x_i) + (1 - t_i)/\left(1 - \hat{e}(x_i)\right)$ -- classical IPW

> Alternative approach: Stage 2: Learn structural model for $\mu(t, x)$
  > Use **weighted** random forest regression with same weights as above

> This is ‘doubly robust’ compared with IPW:
  > Errors in structural model reduced through use of selection-propensity adjustment
  > “ in selection propensity “ “ structural model X-adjustment

Note. This is in the same spirit but not *the* doubly robust estimator by Robins et al. You can find that in Econ-ML.

EARLY APPROACHES: S- and T-LEARNERS

Random forest regression

- S: Learn single model from **all available data**

$$\mu(t, x) = E[y_i | t_i = t, x_i = x]$$

- Then estimate

$$\text{CATE}(x_i) = \mu(1, x_i) - \mu(0, x_i),$$

- T: Learn two models, one for treated, one for untreated:
  - From treated units, learn $\mu_1(x_i) = E[y_i | t_i = 1, x_i]$
  - From control units, learn $\mu_0(x_i) = E[y_i | t_i = 0, x_i]$
  - Combine: $\text{CATE}(x_i) = \mu_1(x_i) - \mu_0(x_i)$

META-ALGORITHMS: X-LEARNER

- Today's focus
- Simply combine learner predictions with observed data:
- As with T-learner:
  - Learn $\mu_1(x_i) = E[y_i | t_i = 1, x_i]$ (from treated units)
  - Learn $\mu_0(x_i) = E[y_i | t_i = 0, x_i]$ (from untreated units)
- 'Impute' individual treatment effects

$$D_i = \begin{cases} D_i^1 \coloneqq y_i - \hat{\mu}_0(x_i) & \text{if } i \text{ is treated} \\ D_i^0 \coloneqq \hat{\mu}_1(x_i) - y_i & \text{if } i \text{ is untreated} \end{cases}$$

Learn $\hat{\tau}_0(x) = E\left[D_i^0 \middle| x_i = x\right]$ and $\hat{\tau}_1(x) = E\left[D_i^1 \middle| x_i = x\right]$

Calculate $\hat{\tau}(x) = \hat{\tau}_0(x)\big(1 - e(x)\big) + \hat{\tau}_1(x)e(x)$

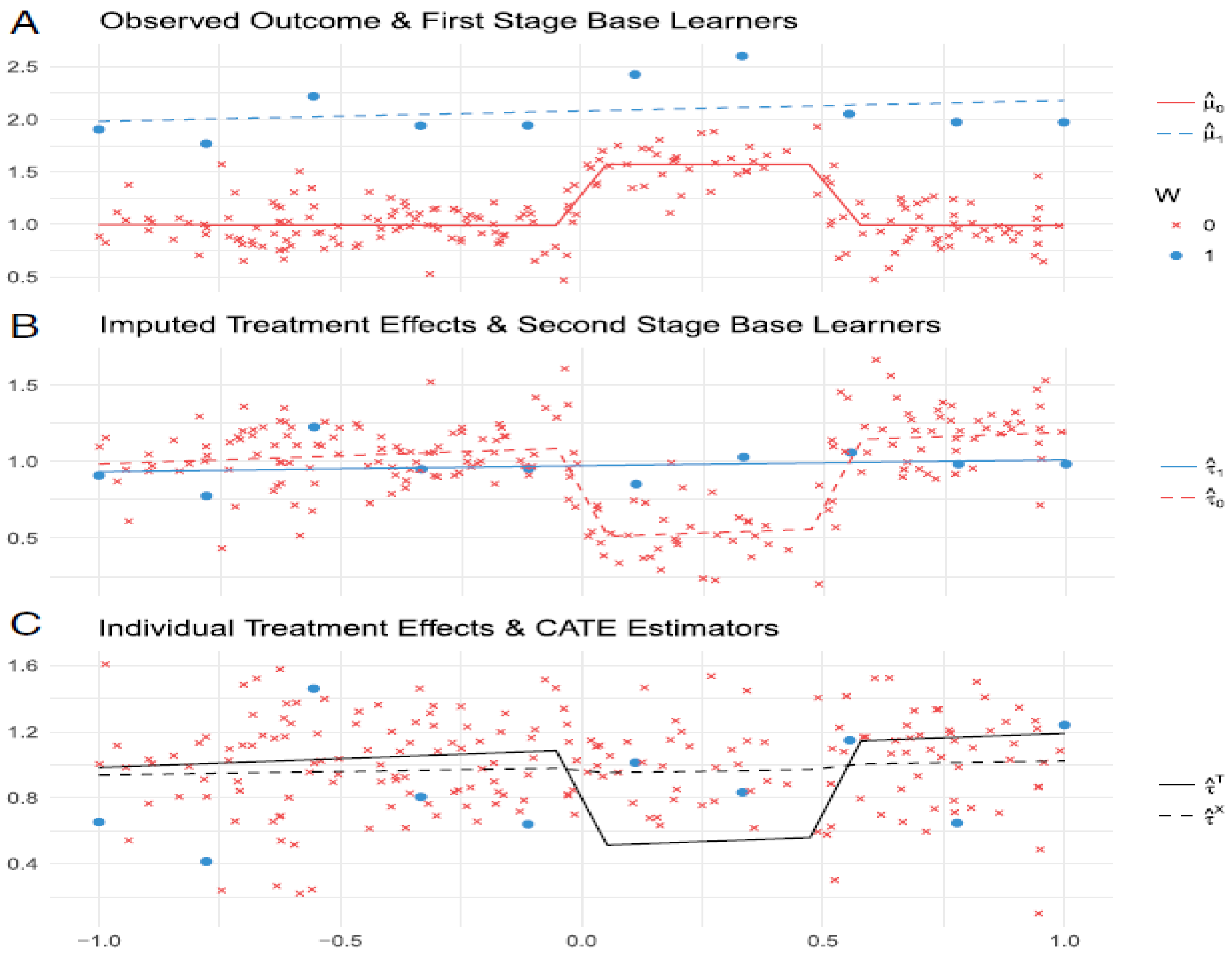


Fig. 1. Intuition behind the X-learner with an unbalanced design. (*A*) Observed outcome and first-stage base learners. (*B*) Imputed treatment effects and second-stage base learners. (*C*) ITEs and CATE estimators.

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…?