Building a Robot Judge: Data Science for Decision-Making

5. Machine Learning and Causal Inference

Learning Objectives

1. Implement and evaluate machine learning pipelines.
2. **Implement and evaluate causal inference designs.**
3. Understand how (not) to use data science tools (ML and CI) to support expert decision-making.

Machine Learning vs Causal Inference

Machine Learning (Weeks 2, 4, and 6):

- in ML, we already know the truth from the dataset.
- we take the labels as given, we just want to predict them.
- we can always verify our model works using the test set.

Causal Inference (Weeks 3, 5, and 7):

- Causal inference is about what *we don't know yet*.
- how do we know if a new policy will work?
  - for example, wearing masks and coronavirus spread.
- There isn't a machine learning dataset to train a model on.
  - we cant experimentally force people to wear a mask or not.
- How do we solve that?

- **Glossary for machine learning vs causal inference terms:** `https://bit.ly/ML-Econ-Glossary`.

Zoom Private Chat Activity: Legal Briefs Dataset

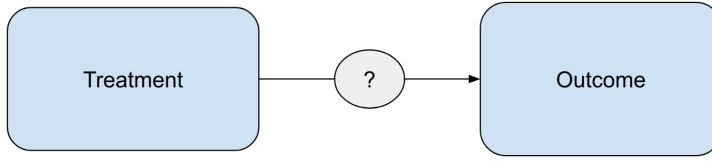
- Let's say we have a dataset of legal briefs with text and metadata, everything you would expect from a case, including information on the actors (litigants, attorneys, judges) and the associated outcomes (e.g. who wins).
- Chat to me privately on zoom:
  1. A **machine learning task or question** that could be addressed with this dataset.
  2. A **causal inference task or question** that could be addressed with this dataset.
- See padlet link in zoom chat – post your answers there.
  - Read your classmates' answers – and "like" them liberally.

Causal inference is needed to improve the world

Consider another critically important policy question:

- In light of coronavirus, should schools reopen or not for in-person teaching?
  - No matter how much we know from lab experiments about the biology/epidemiology of the virus, there will be too much uncertainty about costs/benefits to answer this.
  - We need real-world evidence, but we can't experimentally force schools to reopen or not.
- Can use a natural experiment to produce causal estimates:
  - e.g., variation in number of coronavirus cases before/after openings, using differences in the timing of openings (differences-in-differences).
- Google/Facebook understand this with A/B testing; social scientists want to use this to assist public policy.

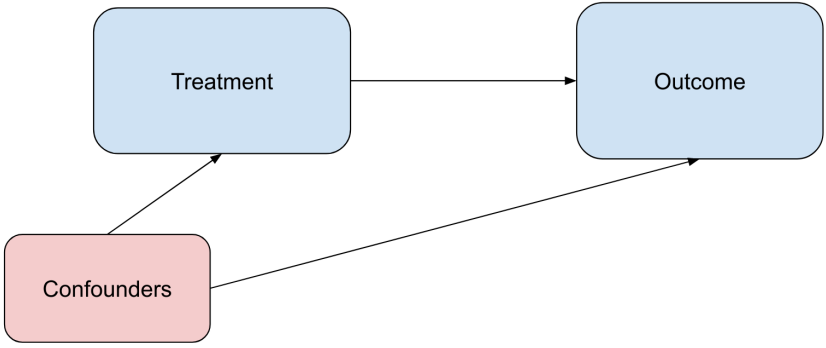
Review: Causal Graphs



- We are interested in estimating a causal effect (if any) of a "treatment" on an "outcome".

Confounders

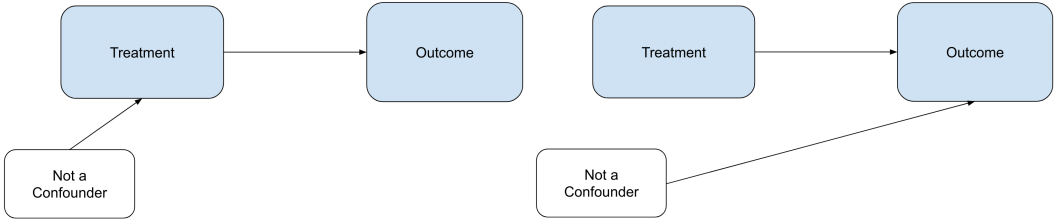
- Confounders affect both the treatment and the outcome:



- Adjusting for confounders (including them in a regression) will reduce bias for estimates of the causal effect of treatment on outcome.
  - Example: ice cream causes heat stroke.

Not Confounders

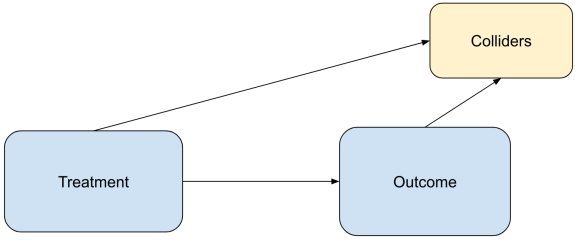
- Variables that affect just the treatment, or just the outcome, are not confounders.

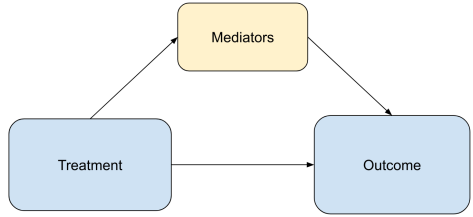


- Adjusting for these variables does not reduce bias, but it might shrink standard errors.

Colliders and Mediators

- **Colliders** are affected by both the treatment and the outcome.
  Mediators are intermediate outcomes / mechanisms.





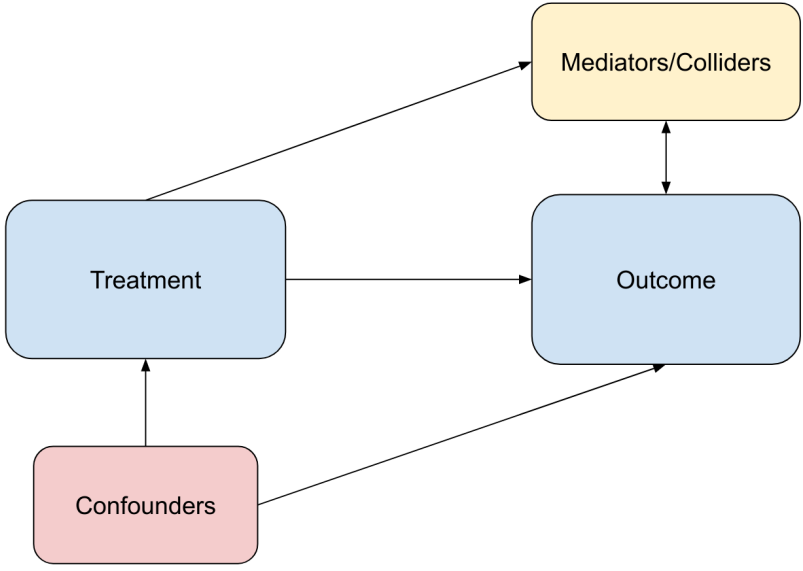
- Adjusting for colliders or mediators will add bias.

Reverse Causation or Joint Causation

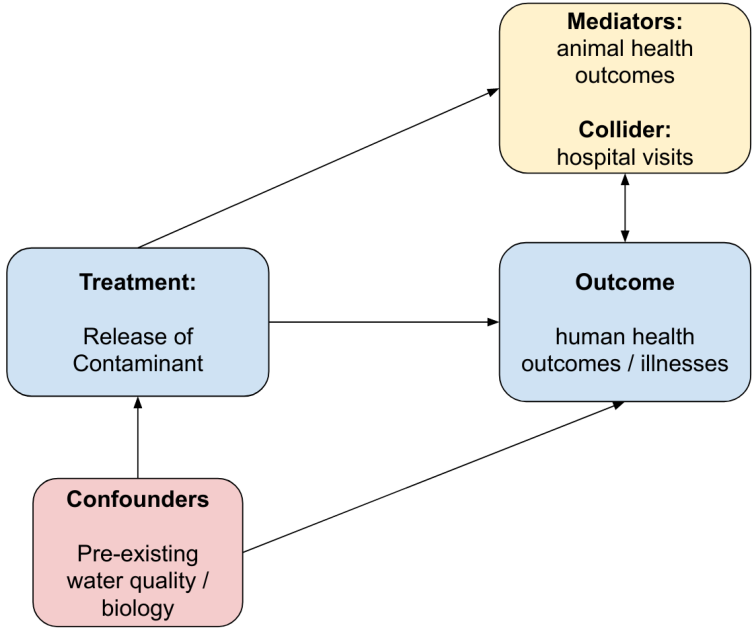
- **Reverse causation**: "the outcome" affects "the "treatment".
  Joint causation: there is bidirectional causation.

- e.g., effect of tax revenues on economic growth.
- Resulting causal estimates are biased, but not adjustable through a confounding variable.
  - Can only solve through a randomized experiment, or through instrumental variables (next week).

Causal Graphs: Overview



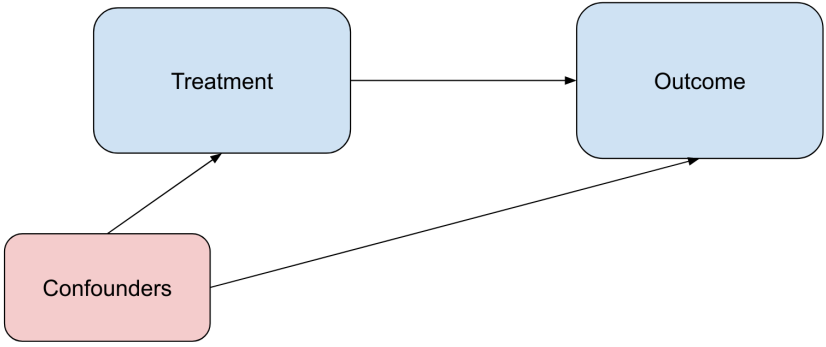
Causal Graph Example: Pollution of a River



Activity: Practice with Causal Graphs

- Think of an example causal inference question:
  - a research question from your field
  - a policy you are interested in
  - a mystery you are fascinated by
- Link to causal graph template posted in zoom chat:
  - make a copy, fill it in
  - make public and paste link into padlet (`bit.ly/BRJ-W5A2`).
  - will review these after the break.

Adjusting (controlling) for observables



- If the treated group and comparison group differ only by a set of observable characteristics, we can "control" or "adjust" for these variables to obtain causal estimates.
- But what if we have 1000 covariates?
  - Machine learning can help.

Propensity Score Matching (PSM)

Rosenbaum and Rubin (1983)

In the case of a binary treatment $D \in \{0,1\}$, the following is equivalent to perfectly adjusting for all observed confounders:

- Predict a one-dimensional "propensity score" $\hat{D}(X) = \Pr(D = 1|X)$, the probability of treatment.
  - can be done with any machine learning model.
- Then include $\hat{D}(X)$ in the regression as a control.
  - best practice: include fixed effects for small bins of $\hat{D}(X)$
  - then all individuals are compared to other individuals with a similar propensity score.
- Caveat: PSM is not optimal in terms of efficiency – standard errors can be reduced by including X rather than $\hat{D}(X)$.

Selecting Controls with Double Lasso (1)

- Consider outcome variable Y and treatment variable D. We want to estimate β from

$$Y = \beta D + g(X) + \epsilon$$

  $g(X)$ is a **"nuisance function"** summarizing the effect of all the confounders.
  - X is a high-dimensional set of predictors – some are confounders, most are not.
  - we will use **lasso** to select which predictors to include in our OLS regression.

Selecting Controls with Double Lasso (2)

- Data prep:
  - drop from X any potential colliders.
  - can add interactions and transformations, e.g. $x_7 x_9$, x_7^2.
  - standardize each variable in X to variance one
- Train two lasso models, $Y \sim \text{Lasso}(X)$ and $D \sim \text{Lasso}(X)$:
  1. use CV grid search across the whole dataset to select best penalties λ_Y and λ_D.
  2. Run both lasso models with whole dataset, get subsets of non-zero predictors, X_Y and X_D
- Regress

$$Y = \beta D + X'_{YD}\gamma + \epsilon$$

where $X_{YD} = X_Y \cup X_D$ is the union of the lasso-selected covariates.

- this is the optimal set of covariates for confounder adjustment (Belloni et al 2014).

Zoom Poll: Which line(s) have a problem?

```
# python
param_grid = {'alpha':  [.01, .1, 1, 10]}
lasso = Lasso()
grid = GridSearchCV(lasso, param_grid)

01 grid.fit(X, Y)
02 alpha_Y = grid.best_params_['alpha']
03 lasso_Y = Lasso(alpha=alpha_Y)
04 lasso_Y.fit(X,Y)
05 selected_Y = lasso_Y.coef_ != 0

06 grid.fit(X, D)
07 alpha_D = grid.best_params_['alpha']
08 lasso_D = Lasso(alpha=alpha_D)
09 lasso_D.fit(X,D)
10 selected_D = lasso_D.coef_ != 0

11 X_YD = X[:,(selected_Y and selected_D)]

* stata (X_YD_1, X_YD_2, ...  are selected covariates)
12 reghd Y D X_YD_*, robust
```

What if $g(\cdot)$ is not linear?

- Lasso assumes that $g(X)$ is linear in X.
  - we somewhat relaxed that assumption by adding interactions and quadratic transformations.
  - but how do we know what interactions/transformations to add?
- Can use a non-linear model, e.g. xgboost, to approximate and adjust for $g(X)$.
  $\rightarrow$ Double Machine Learning, also called Debiased machine Learning (Chernozhukov et al 2018)

Double ML: Setup

$$Y = \beta D + g(X) + \epsilon$$

- low-dimensional treatment D, high-dimensional set of (observed) confounders X.
  - OLS regression without adjusting for confounders will be biased for $\hat{\beta}$
  - can we just include them in the regression as linear covariates?
    - will not adjust correctly due to potential non-linearities.
    - will probably fail to converge due to high dimensionality / collinearity / overfitting

Double ML method

1. Learn Y given X, $\hat{Y}(X)$, using any ML method
2. Learn D given X, $\hat{D}(X)$, using any ML method
3. Form residuals $\tilde{Y} = Y - \hat{Y}(X)$ and $\tilde{D} = D - \hat{D}(X)$
4. Regress $\tilde{Y}$ on $\tilde{D}$ to learn $\hat{\beta}$.

Cross-Fitting: Split into samples A and B, 50% of data each, to prevent overfitting:

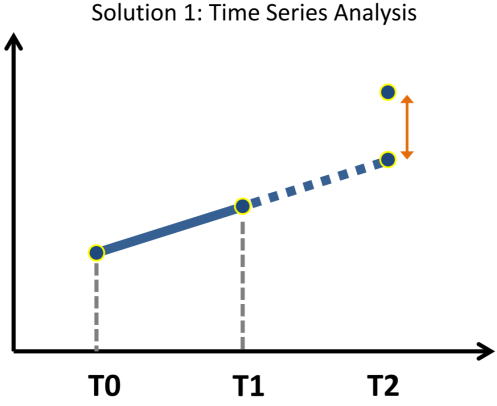
- Fit (1) and (2) on sample A, then predict (3) and regress (4) on sample B, to estimate $\hat{\beta}_A$
- vice versa: fit (1)/(2) on sample B, and predict/regress (3)/(4) on sample A, to learn a second estimate for $\hat{\beta}_B$.
- average them to get a more efficient estimator:$\hat{\beta} = \frac{1}{2}(\hat{\beta}_A + \hat{\beta}_B)$.

Recap: Panel Data

Panel data, or longitudinal data, is data over time.

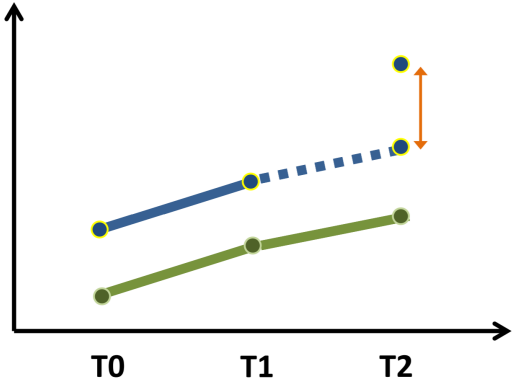
- our goal is to use this data to construct counterfactuals and learn causal effects, for example due to changes in government policies.

Time Series Analysis



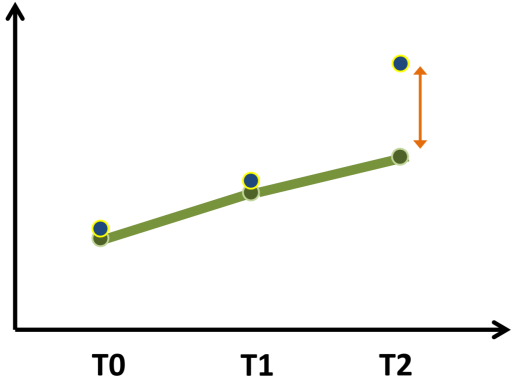
- In macroeconomics (analysis of the whole economy), you only observe one individual/group (the economy).
  - How to estimates causal effect of a macroeconomic policy like changing interest rates?
- "time series analysis" assumes economy continues on current trend in absence of intervention.
  - macro is wrong most of the time.

Differences-in-Differences (last week)



- Differences-in-differences (or fixed effects) estimation assumes that the treated group (in blue) would evolve the same way as the control group (in green)

Matching



- Matching searches over the control group and finds an individual with similar pre-trends.
  - can also match on covariates
  - can use any distance metric, e.g. cosine similarity.
- Esimate causal effect as the difference between the treated unit and the matched unit.

Synthetic Control

- **synthetic control**: construct a synthetic "match" from a weighted average of other individuals (based on covariates).
- Statistically comparable to fixed effects or matching, but **powered up with ML.**

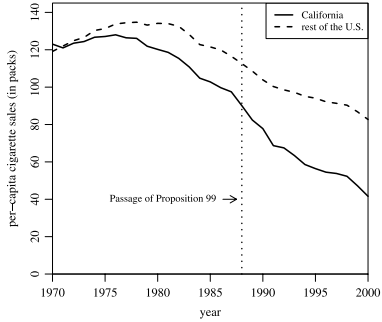
Example: Tobacco Laws in CA

In 1988, California passed anti-tobacco legislation (Proposition 99)

- Increased tax by \$0.25/pack
- Extra tax revenues earmarked to health budget
- Funded anti-smoking campaigns
- Clean-air signs in closed spaces

```
// stata
use http://fmwww.bc.edu/repec/bocode/s/synth_smoking.dta
```

Trends in cigarette sales: Parallel trends fails



Source: Abadie, Diamond and Hainmueller (2010)

Rest of US is not a good comparison group for California

- Trends start diverging in 1970s, before the reform
- Parallel trends assumption fails $\Rightarrow$ Cannot apply diff-in-diff

Synthetic Control Setup

- *Dataset:*
  - $j \in \{1, ...J+1\}$ *units,* $t = 1, 2, ..., T$ *periods.*
  - Outcome Y_{jt} (e.g. smoking), characteristics X_j
- Treatment:
  - Unit 1 (e.g. California) is exposed to intervention in periods $T_0 + 1, ..., T$
- Control group:
  - Remaining J units (other states) are potential controls ("donor pool")
- Objective:
  - find combination of untreated units that best approximates treated unit

Formalization

- Define weights $w_2, ..., w_{J+1} \geq 0$ where $\sum_{i=2}^{J+1} w_i = 1$.

$$\text{Synthetic Control Treatment Effect} = \underbrace{Y_{1t}}_{\text{treated}} - \underbrace{\sum_{j=2}^{J+1} w_j^* Y_{jt}}_{\text{synthetic}}$$

  where $t \geq T_0$ is the post-intervention period

- Synthetic control vector $\left(w_2^*, ..., w_{J+1}^*\right)$ is chosen to minimize

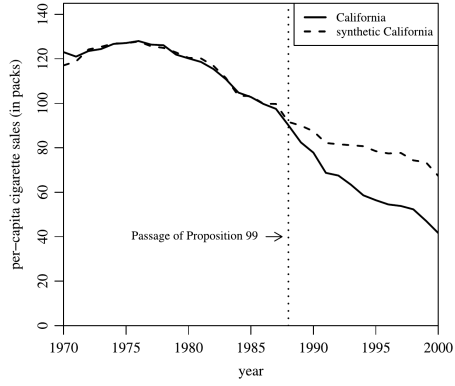
$$\sum_{m=1}^{k} v_m \left(x_{1m} - \sum_{j=2}^{J+1} w_j x_{jm} \right)^2$$

  - v_m = weight on m-th variable, chosen to minimize pre-reform MSE for Y, that is, match on pre-trends.

Synthetic California

Table 2. State weights in the synthetic California

State	Weight	State	Weight
Alabama	0	Montana	0.199
Alaska	–	Nebraska	0
Arizona	–	Nevada	0.234
Arkansas	0	New Hampshire	0
Colorado	0.164	New Jersey	–
Connecticut	0.069	New Mexico	0
Delaware	0	New York	–
District of Columbia	–	North Carolina	0
Florida	–	North Dakota	0
Georgia	0	Ohio	0
Hawaii	–	Oklahoma	0
Idaho	0	Oregon	–
Illinois	0	Pennsylvania	0
Indiana	0	Rhode Island	0
Iowa	0	South Carolina	0
Kansas	0	South Dakota	0
Kentucky	0	Tennessee	0
Louisiana	0	Texas	0
Maine	0	Utah	0.334
Maryland	–	Vermont	0
Massachusetts	–	Virginia	0
Michigan	–	Washington	–
Minnesota	0	West Virginia	0
Mississippi	0	Wisconsin	0
Missouri	0	Wyoming	0

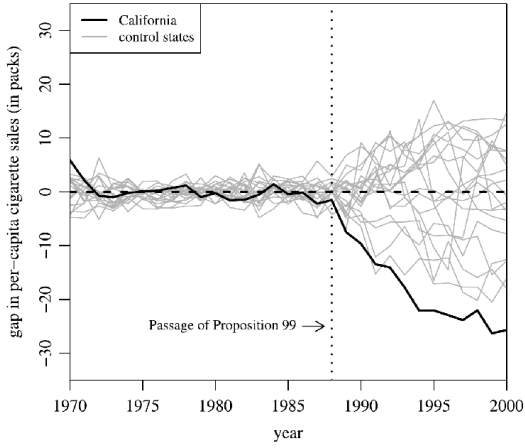


Source: Abadie, Diamond and Hainmueller (2010)

```
// stata
tsset state year
synth y x1 x2 x3 y(1980) y(1988), trunit(3) trperiod(1989)
// trunit = 3 is the state # for california
// trperiod = 1989 is the first post-reform year
```

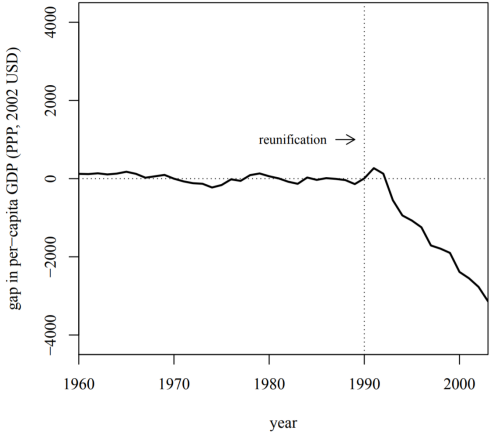
Inference

- Synthetic control does not give standard errors. Instead, use bootstrap approach:
  - Compare estimated synthetic control effect for california to distribution of placebo effects where treated unit is picked at random from donor pool.



Application 2: Effect of Reunification on West Germany GDP

Country	Weight	Country	Weight
Australia	0	Netherlands	0.11
Austria	0.47	New Zealand	0.11
Belgium	0	Norway	0
Canada	0	Portugal	0
Denmark	0	Spain	0
France	0	Sweden	0
Greece	0	Switzerland	0
Ireland	0	United Kingdom	0.17
Italy	0	United States	0
Japan	0		0.14



Practicalities

- Easiest way to get going with synthetic control: the synth package in stata.
- In python, can use microsoft's SparseSC package.

```
# python
from numpy import hstack
from SparseSC import fit

# Let X be the features plus some of the targets
X = hstack([features, targets[:,:t])

# And let Y be the remaining targets
Y = targets[:,t:]

# fit the model:
sc = fit(X=X, Y=Y, model_type="full")
```

Summary: Synthetic Control

Advantages:

1. Works with a single treated unit.
2. Makes explicit the contribution of each comparison unit to the synthetic control
3. Quantitative and qualitative ways to analyze similarities and differences of treatment and synthetic control
4. Formalizing how comparison units are chosen has nice properties for inference

Limitations:

1. Still requires parallel trends / counterfactual assumption.
2. Could be idiosyncratic shocks to treated unit or comparison units
3. Cannot estimate effect of a single reform when multiple reforms passed at once

Heterogeneous Treatment Effects

- Treatments don't affect every individual equally.
  - for example, effect of covid social distancing will depend on the age distribution.
- The simplest way to estimate these is to interact treatment with another covariate (the "**moderator**"):

$$Y_i = \beta_1 D_i + \beta_2 \text{Age}_i + \beta_3 D_i \text{Age}_i + \epsilon_i$$

  - here, β_3 summarizes heterogeneous impact by age ($\frac{\partial Y_i}{\partial D_i} = \beta_1 + \beta_3 \text{Age}_i$)

Activity: Brainstorming about Moderators

Revisit your customized causal graph:

- Add a new "bubble" with the header "Moderators", and list some potential variables/characteristics that you expect to have a larger or smaller treatment effect.

Conditional Treatment Effects

- Consider the model

$$Y = \underbrace{\beta(X)}_{\text{CTE}} D + g(X) + \epsilon$$

  - the causal estimate $\beta(X)$ is a function of X.
  - this is the conditional treatment effect (CTE) – that is, the effect conditional on X.
- Can learn flexible representation of $\hat{\beta}(X)$ using machine learning.

T-Learner Method

- Residualize Y on the fixed effects and controls to get rid of $g(X)$:

$$Y = \beta(X)D + \epsilon$$

  - if D is randomly assigned (e.g. RCT), this is not necessary.
  - Note that the standard (non-conditional) treatment effect is then immediately obtainable by OLS: $\hat{\beta}_{OLS} = \text{Cov}(Y, D)/\text{Var}(D)$.

T-Learner Method:

- Using any machine learning method:
  - Learn $\mu_0(X) = \mathbb{E}(Y|X, D = 0)$
  - Learn $\mu_1(X) = \mathbb{E}(Y|X, D = 1)$
- Tune parameters in whole dataset using cross-validation.
- The conditional treatment effect estimate is $\hat{\beta}(X) = \mu_1(X) - \mu_0(X)$.

See Knaus, Lechner, and Strittmatter (2020) for a review of different methods/extensions.

Homework Note

- Next homework is due Tuesday October 27th before 8am.

Two parts:

1. Complete a jupyter notebook on xgboost and double ML.
   - Reminder: Don't submit an IPYNB file. please export them to HTML or PDF before submitting.
2. Peer review of response essays:
   - You will be randomly assigned two anonymized essays from one of your classmates.
   - Write one paragraph (5-10 sentences) about each essay, providing constructive feedback/suggestions.
   - Follow the rubric on the homework assignments page.