ECON 520: Data Science for Economists

Introduction to Causal Inference

Pedro H. C. Sant'Anna

EMORY UNIVERSITY

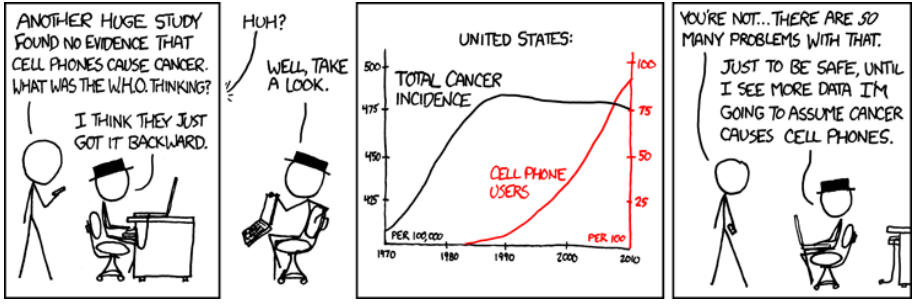
Spring 2024

Recap

- So far, in this course we have spent a lot of time dedicated to the “Data Science” pipeline.
- Focus has been on tools:
  - Github for version control
  - Writing clean code
  - Web scrapping
  - Data wrangling
  - Data visualization
- Today, we will start focusing on the **Causality Mindset**!

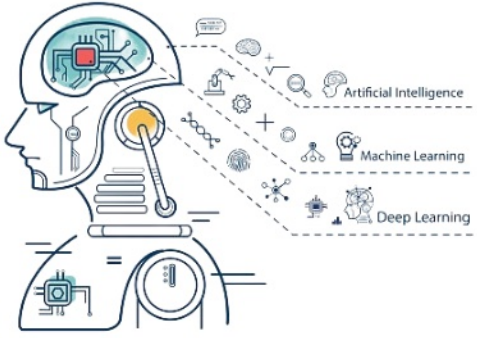
Correlation and causation

Correlation and causation



Source: https://xkcd.com/925/

Maybe machine learning can help us



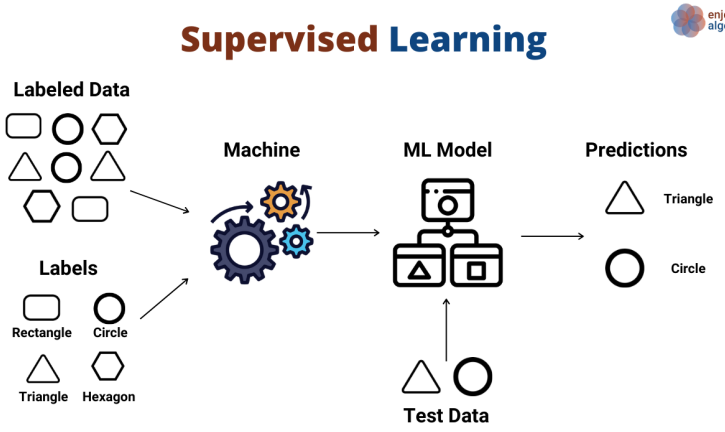
Supervised Machine Learning: Examples

- Machine learning has been featured in many aspects of our lives.
  - Recommendation systems
  - Self-driving cars
  - Drones
  - Auto-translation
  - Image recognition

Supervised Machine Learning as Prediction

- Common theme of these applications: all types of “prediction” problems.
  - We have data on features (predictors) and labels (outcomes)
  - Learn (extract/estimate) what matters to make a good prediction (stable environment)
  - We can assess if the prediction was right or wrong
  - **The goal of supervised learning is to get the prediction right!**

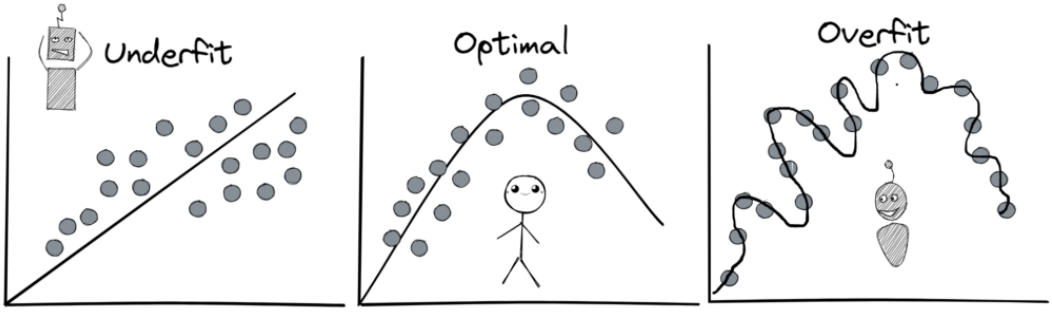
Supervised Learning



Source: enjoy algorithms

Overfittting

Overfittting



We are interested in topics beyond prediction, too

- Prediction and Forecasting are important topics for us, Economists and Data Scientist!
- But we are also interested in **"What-if"** types of questions that do not fit well into this prediction framework.
  - For Amazon, what is the impact of having a product out-of-stock on long-term free cash flow?
  - For Microsoft, what is the impact of Dall-E and Co-Pilot on Bing usage?
  - For Etsy, what is the impact of displaying product ratings on the number of orders?
  - For Health Care providers, what is the incremental impact of drug B vs. drug A on the time-to-recovery of a sick patient?
  - **Other Examples?**

Causality definitions from philosophers

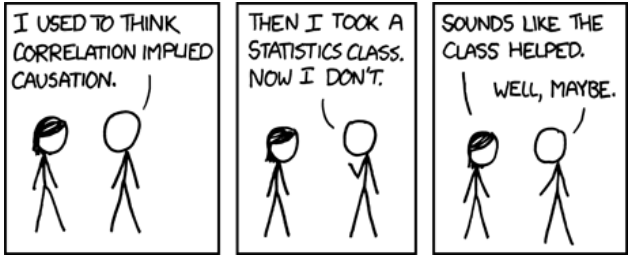
"If a person eats of a particular dish, and dies in consequence, that is, would not have died if he had not eaten it, people would be apt to say that eating of that dish was the source of his death." – John Stuart Mill (19th-century moral philosopher and economist)

"Causation is something that makes a difference, and the difference it makes must be a difference from what would have happened without it." – David Lewis (20th-century philosopher)

Causal Inference is hard

- Mill's counterfactuals were immensely valuable for the clarity of the definition as well as its intuitive validity of causality.
- But it also made it clear that causality is a tricky business!
- If I have to know what would have happened had I not eaten the dish, but I did eat the dish, **how would I ever be able to know the causal effect of eating the dish?**
- The same reasoning applies to all the what-if types of questions we have discussed.
- This is a valid concern, but this should not stop us from being able to ask questions!

Causality is a game of counterfactuals



Source: https://xkcd.com/552/

Off-the-shelf application of supervised learning is not great

- Causal inference is not a prediction problem but rather a counterfactual problem.
- This makes things challenging because:
  - Direct use of ML methods is biased for causal effect due to confounding.
  - ML aims at minimizing prediction error, not counterfactuals.
  - We never observed the true causal effect, which makes model selection trickier.
  - We are not only interested in the counterfactual itself but also in quantifying its uncertainty (making inferences).
- Some modifications and tricks can be used to bypass several of these.
- **Key: decompose the problem into predictive and causal parts.**

The approach we adopt in causal inference

1. Specify the causal question of interest and map that into a causal target parameter.

2. Figure out a research design using domain knowledge that can credibly answer the question (**We usually call this identification)**

    - We usually abstract from sample size considerations here.

    - What we want is the “right” data that leverages a “quasi-random” variation in our treatment variable.

Approach we adopt in causal inference

3. State our assumptions, and provide supportive evidence of their credibility in our context.
    - Why treatment is “quasi-random”: A/B test is actually effective!
    - For which population you can identify the effects
4. Pick an estimation and inference method that is effective and reliable.
    - This is where ML will come into play: random forest, neural nets, boosting, GAN, among other methods
    - How to tune (or regularize) these methods to give credible results.

Examples of motivating causal questions

- What is the “penalty” impact of having a product out-of-stock on long-term free cash flow?
- What is the impact of displaying product ratings on the number of orders?
- What is the impact of following a sale recommendation on Gross Merchandise Sales (GMS)?
- What is the impact of offering Game Pass rewards on engagement among those who play? What about among those who received the rewards offered?
- What is the effect of claiming Game Pass rewards on engagement among claimers?
- What is the effect of being offered a job training on earnings? What about the effect of receiving the job training on future earnings?

Notation leveraging potential outcomes

- We will adopt the **Rubin Causal Model** and define potential outcomes.

  There are other approaches/languages out there, too, e.g., Judea Pearl's Directed Acyclic Graph (DAG). They should be seen as complements.

- Potential outcomes define outcomes in different states of the world, depending on the type of treatment units assigned to them.

- Let D be a treatment variable.
  - When D is binary, $D_i = 1$ means unit i is treated, and $D_i = 0$ means unit i is not treated.
  - When D is multi-valued, $D \in \{0, 1, 2, \ldots, K\}$, $D_i = d$ means unit i received treatment d.
  - When D is continuous, $D \in [a, b]$, $D_i = d$ means unit i received treatment d.

- Let $Y_i(d)$ be the potential outcome for unit i if they were assigned treatment d.

- Each unit i has a lot of **different** potential outcomes

Notation based on application about sale recommendation

- What is the impact of following the recommendation to “create a sale” on Gross Merchandise Sales (GMS)?

  - **Treatment $\mathbf{D}$:**
    $D_i = 1$ if the seller creates a sale following the sale recommendation for product i.
    $D_i = 0$ if the seller does not create a sale ignoring the recommendation.

  - **Potential Outcomes $\mathbf{Y_i(1)}$, $\mathbf{Y_i(0)}$**
    $Y_i(1)$ GMS of product i if the seller were to follow the create a sale recommendation.
    $Y_i(0)$ GMS of product i, if the seller were not to follow the create a sale recommendation

Causality with potential outcomes

- **Unit-specific Treatment Effect**
  - The treatment effect or causal effect of switching treatment from d' to d is the difference between these two potential outcomes:

$$Y_i(d) - Y_i(d')$$

  - When treatment is binary,

$$Y_i(1) - Y_i(0)$$

Fundamental problem of causal inference

- **Fundamental problem of causal inference**:

  For each unit i, we cannot observe their different potential outcomes at the same time. We only see one of them.

- **Observed outcome with binary treatments**

  - Observed outcomes for unit i are realized as

$$Y_i = 1\{D_i = 1\}Y_i(1) + 1\{D_i = 0\}Y_i(0)$$

$$Y_i = \begin{cases} Y_i(1) \text{ if } D_i = 1 \\ \\ Y_i(0) \text{ if } D_i = 0 \end{cases}$$

Fundamental problem of causal inference: Missing data

	Data				
Unit	$Y_i(1)$	$Y_i(0)$	D_i	$Y_i(1) - Y_i(0)$	X_i
1	?	✓	0	?	x_1
2	✓	?	1	?	x_2
3	?	✓	0	?	x_3
4	✓	?	1	?	
⋮	⋮	⋮	⋮	⋮	⋮
n	✓	?	1	?	x_n

✓: Observed data
?: Missing data (unobserved counterfactuals)

Causality with potential outcomes

- **Problem:**
  - Causal inference is difficult because it involves missing data.
  - How can we find $Y_i(1) - Y_i(0)$?

- **"Cheap" solution - Rule out heterogeneity.**
  - $Y_i(1), Y_i(0)$ constant across units.
  - Assuming all potential outcomes are the same is **very strong**: who believes in that?!

Very little hope for learning about unit-specific treatment effects.

We will acknowledge that learning unit-specific TEs is hard, if not impossible.

We will focus on treatment effects in an average sense, but allow them to vary with X.

For simplicity, we will focus on binary treatment setups.

Parameters of interest: Level metrics

- **ATT**: The Average Treatment Effect among the Treated units is

$$ATT = \mathbb{E}\left[Y_i(1) - Y_i(0) | D_i = 1\right]$$

What is the effect of following the “create-a-sale” recommendation on GMS, among products that actually created the sale?

Particularly useful to assess if products that followed the recommendation actually benefit from it.

- **ATU**: The Average Treatment Effect among Untreated units is

$$ATU = \mathbb{E}\left[Y_i(1) - Y_i(0) | D_i = 0\right]$$

What is the effect of following the “create-a-sale” recommendation on GMS, among products that received the recommendation but did not create created the sale?

Particularly useful to assess if products that received but did followed the recommendation would have benefited from it.

Parameters of interest: Level metrics

- **ATE**: The (overall) Average Treatment Effect is

$$ATE = \mathbb{E}\left[Y_i(1) - Y_i(0)\right]$$

What is the effect of following the “create-a-sale” recommendation on GMS, among products that received the recommendation?

Particularly useful to assess the value of recommendations if they were to be followed automatically.

What if we want to express

average causal effects

as relative lifts?

Parameters of interest: Relative metrics

- All the average causal parameters discussed so far are expressed in the same units as Y.
- If Y is expressed in dollars, ATE, ATT and ATU will also be expressed in dollars.
- If Y is expressed in number of units shipped, ATE, ATT and ATU will also be expressed in number of units shipped.
- Sometimes, want to translate the ATE, ATT or ATU into percentage terms.

Parameters of interest: Level metrics

- **RATT**: The Relative Average Treatment Effect among the Treated units is

$$RATT = \frac{\mathbb{E}\left[Y_i(1) - Y_i(0) | D_i = 1\right]}{\mathbb{E}\left[Y_i(0) | D_i = 1\right]}$$

- **RATU**: The Relative Average Treatment Effect among Untreated units is

$$RATU = \frac{\mathbb{E}\left[Y_i(1) - Y_i(0) | D_i = 0\right]}{\mathbb{E}\left[Y_i(0) | D_i = 0\right]}$$

- **RATE**: The Relative Average Treatment Effect is

$$RATE = \frac{\mathbb{E}\left[Y_i(1) - Y_i(0)\right]}{\mathbb{E}\left[Y_i(0)\right]}$$

What if we want to understand how the average causal effects vary with observable features?

Parameters of interest: Conditional Average Treatment Effects

Let X_{all} be a set of features/covariates available to you, and let X_s be a subset of X_{all}.

- **CATE**: The Conditional Average Treatment Effect given X_s is

$$CATE_{X_s}(x_s) = \mathbb{E}\left[Y_i(1) - Y_i(0) | X_s = x_s\right]$$

How does the effect of following the "create-a-sale" recommendation on GMS vary with regions and product types?

Other parameters have similar characterizations.

Before moving on, ensure to map your Q to a parameter

- Ideally, we should all acknowledge the types of causal parameters we are after.
- This also requires carefully defining what exactly the treatment is.
- When treatment is multi-valued or continuous, there will be even more parameters of interest.
  - Marginal incremental treatment is one type of intervention.
  - Switching between no treatment to a given treatment level is another type of intervention.

Imputing counterfactuals

- All the causal questions and causal parameters discussed so far involved counterfactual quantities.
- For example, we may need to get $\mathbb{E}[Y(1)]$, $\mathbb{E}[Y(0)]$, $\mathbb{E}[Y(1)|X]$, and $\mathbb{E}[Y(0)|X]$ to answer many of the questions
- We already acknowledge we do not see these potential outcomes running around us.
- What if we ignore this and directly compare treated vs. untreated units?

Selection Bias

Problem:
Comparison of outcomes between the treated and the untreated units do not usually give the right answer.

$$\begin{aligned}
\mathbb{E}\left[Y_i|D_i=1\right]-\mathbb{E}\left[Y_i|D_i=0\right] &= \mathbb{E}\left[Y_i(1)|D_i=1\right]-\mathbb{E}\left[Y_i(0)|D_i=0\right] \\
&= \mathbb{E}\left[Y_i(1)-Y_i(0)|D_i=1\right] \\
&\quad +\left(\mathbb{E}\left[Y_i(0)|D_i=1\right]-\mathbb{E}\left[Y_i(0)|D_i=0\right]\right) \\
&= ATT + \textit{Selection Bias}
\end{aligned}$$

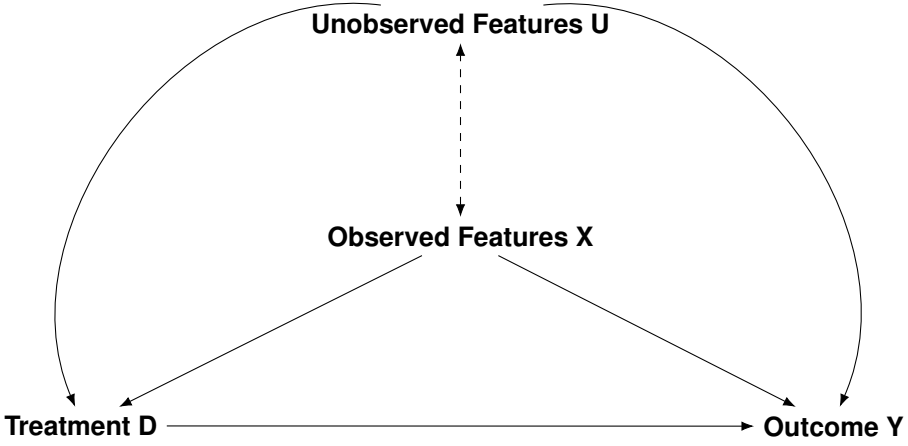
- Selection Bias term **unlikely** to be zero in most applications that do not have a lab.
- Selection into treatment is often associated with the potential outcomes (directly or indirectly).

How does selection bias appear?

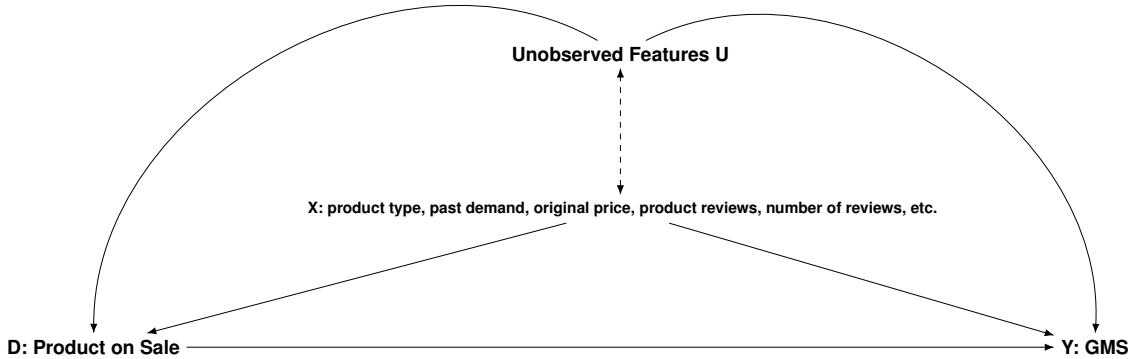
- Selection bias appears because treatment selection is confounded with other factors.

- Treated and untreated units have different observed and unobserved features that can also explain treatment effects.

Selection bias at play: Hardest but realistic case

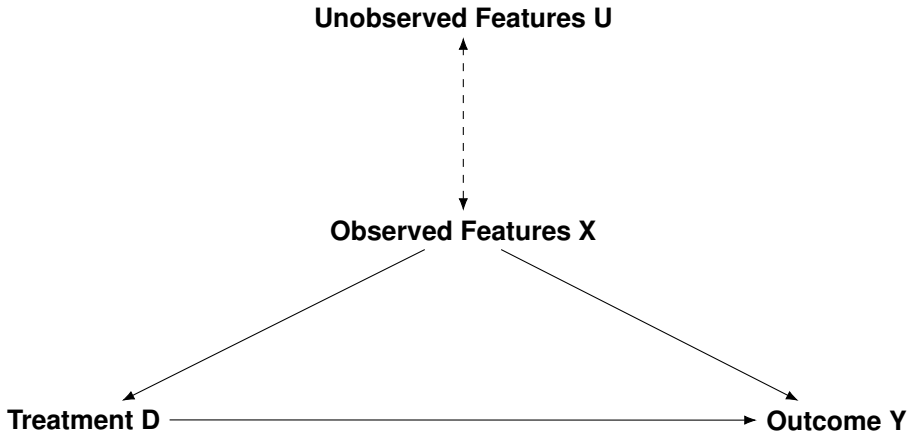


Selection bias at play: Hardest but realistic case

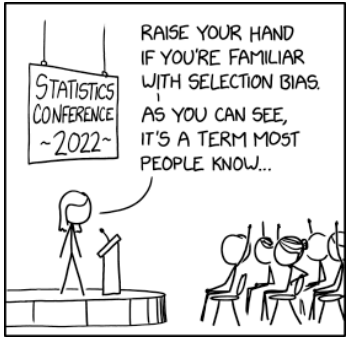


GMS: Gross Merchandise Sales

Selection bias at play: A more hopeful case



Selection Bias



Source: https://xkcd.com/2618/

Selection bias in real life

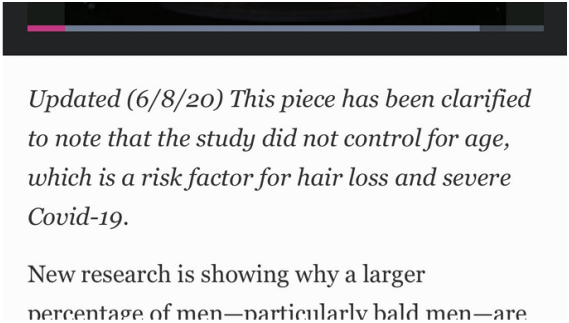
CORONAVIRUS | 130,196 views | Jun 6, 2020, 11:26am EDT

Bald Men At Higher Risk Of Severe Coronavirus Symptoms

Marla Milling Contributor ⓘ

Healthcare

I am a Forbes.com Contributor specializing in geriatric health and women's health articles.



Updated (6/8/20) This piece has been clarified to note that the study did not control for age, which is a risk factor for hair loss and severe Covid-19.

New research is showing why a larger percentage of men—particularly bald men—are

Source:

https://www.forbes.com/sites/marlamilling/2020/06/06/bald-men-at-higher-risk-of-severe-coronavirus-symptoms/

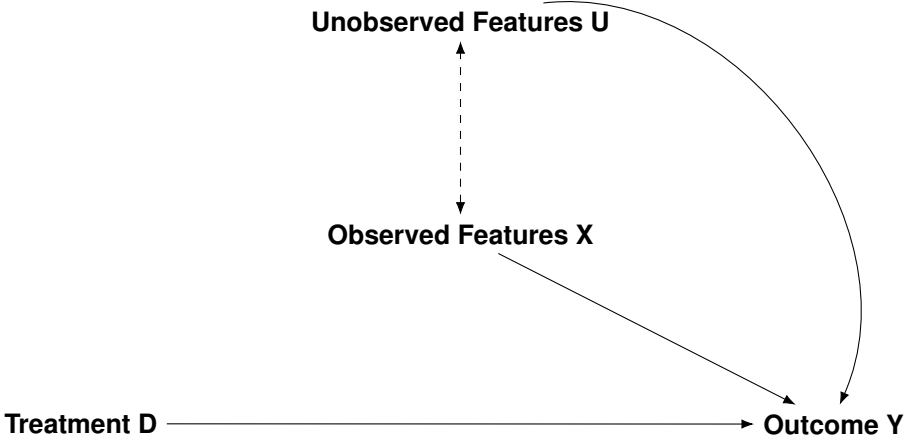
Link to paper

Randomization

- The challenge in Causal inference is that those with $D = 1$ and those with $D = 0$ are not generally comparable.
- But what if we were able to make them comparable?
- How?
- Randomly allocating units to treatment or control group:

$$Y(1), Y(0) \perp\!\!\!\perp D$$

Power of complete randomization



With complete randomization, life is simple again

- With complete randomization, for each treatment level $d \in \mathcal{D}$,

$$\mathbb{E}\left[Y_i | D_i = d\right] = \mathbb{E}\left[Y_i(d) | D_i = d\right] = \mathbb{E}\left[Y_i(d)\right].$$

- Thus,

$$\mathbb{E}\left[Y_i | D_i = 1\right] - \mathbb{E}\left[Y_i | D_i = 0\right] = ATE = ATT = ATU.$$

- We can estimate these using the plug-in principle: replace population $\mathbb{E}\left[\cdot|\cdot\right]$ with their sample analogs, $\mathbb{E}_n[\cdot|\cdot]$

- For level metrics, just run the linear regression

$$Y_i = \alpha + \beta D_i + \varepsilon_i$$

and $\widehat{\beta}$ will give you an estimator for the ATE.

What about understanding treatment effect heterogeneity?

How treatment effects vary across product types?

Treatment Effect Heterogeneity wrt discrete features

- With a successful A/B test, it is relatively easy to understand treatment effect heterogeneity with respect to discrete features.
- All you need to do is subset the data with respect to the features you care about and re-run the analysis multiple times.
- The same is true for understanding how treatment effects vary over time.

Treatment Effect Heterogeneity wrt discrete features

- We can use some tricks here to run this in a single regression (at least with level metrics).
- Suppose there are K different subgroups of interest (think each subgroup is a product type), i.e., $X_s \in \{1, 2, \ldots, K\}$.
- Let $\widetilde{X}_{k,i} = 1\{X_i = k\}$ be a dummy variable if a unit i belongs to subgroup k.
- To assess how the ATEs vary across subgroups, we leverage the following regression:

$$Y_i = \beta D_i + \sum_{k=1}^{K} \left(\gamma_k \widetilde{X}_{k,i} + \theta_k D_i \widetilde{X}_{k,i} \right) + \varepsilon_i$$

- Easy to show that

$$CATE_{X_s}(k) \equiv \mathbb{E}\left[Y(1) - Y(0) | X_s = k\right] = \beta + \theta_k.$$

What if the features are continuous?

- If the features are continuous, we need to do some extra work.
- Here, we could leverage the more flexible regression specification

$$Y_i = \gamma(X_{s,i}) + \theta(X_{s,i})\, D_i + \varepsilon_i$$

- Easy to show that

$$\begin{aligned}
\mathbb{E}\left[Y(0)|X_s\right] &= \gamma(X_{sub,i}), \\
\mathbb{E}\left[Y(1)|X_s\right] &= \gamma(X_{sub,i}) + \theta(X_{sub,i}), \\
CATE_{X_s}(x) &= \theta(x)
\end{aligned}$$

- We can use ML methods to estimate the unknown functions, $\gamma(\cdot)$ and $\theta(\cdot)$

  See, e.g., Athey and Wager (2018), Athey, Tibshirani and Wager (2019), Chernozhukov, Demirer, Duflo and Fernández-Val (2022) and references therein.

We will need assumptions to make this fly now

- Without an A/B test, things are definitely more complicated.
- Now, data by itself will not be enough
- We will need to start owning our assumptions and look for datasets where these are "reasonable"
- Here, perfect is the enemy of doable (at least as a first-order approximation).
- There are many different possible paths in terms of assumptions
- We will cover only unconfoundedness (arguably the most popular in industry).

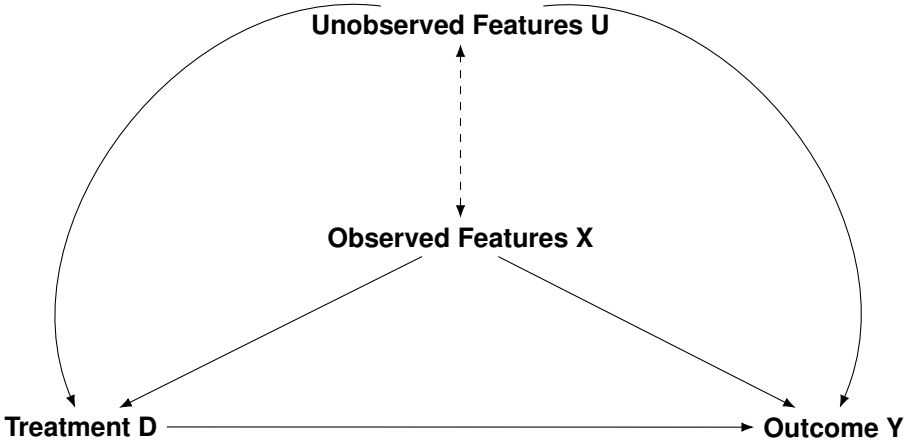
Unconfoundedness

- The first approach to proceed with our causal analysis is to impose unconfoundedness:

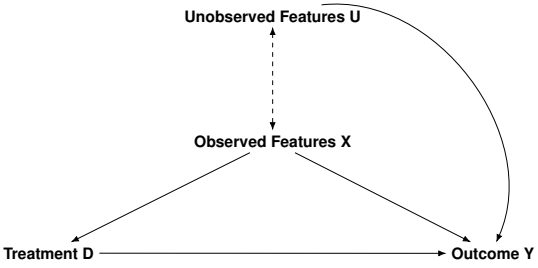
$$Y(1), Y(0) \perp\!\!\!\perp D | \, X_{all}$$

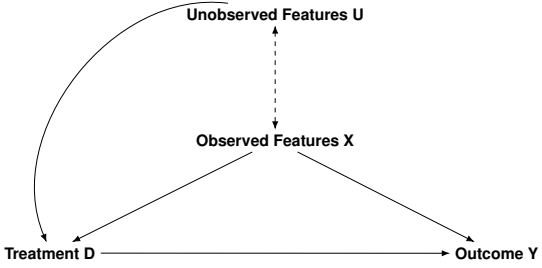
- This assumption states that once we account for observed characteristics X_{all}, selection into treatment is as-good-as-random.
  - This means that everything that affects both treatment D and potential outcomes $Y(1), Y(0)$ is observed and accounted for.
- We will also impose a common support assumption that units within each "strata" defined by X_{all} has a positive probability of being in the treatment and the untreated group.
  - This assumption rules out deterministic treatment allocations

Recall our starting point without randomization

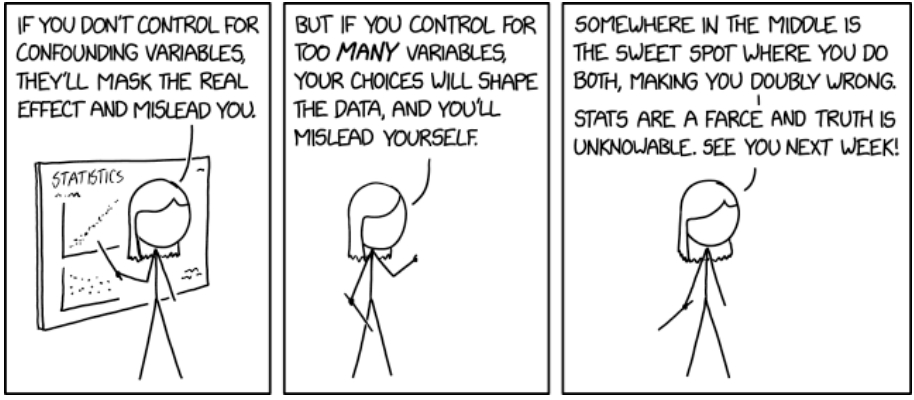


Unconfoundedness imposes that one of these two is true





Confounding Factors: the danger



Source: https://xkcd.com/2560/

Identification of ATE under unconfoundedness

- Under unconfoundedness, we need to explicitly account for confounding factors X_{all}.
- One can use different estimation strategies to account for X_{all}, including regression adjustments and reweighting.
- Intuition for the regression adjustment is the following:
  1. "Stratify" the population based on values of X_{all}
  2. Within each stratum, make a simple comparison of means between treated and untreated units.
  3. Aggregate all the strata-specific ATE and report an estimate of the ATE

Estimating ATE under Unconfoundedness

ATE under unconfoundedness: Regression Adjustments

- Let me give a more mathy explanation
- Let $m_1(X_{all}) = \mathbb{E}[Y|D=1, X_{all}]$ and $m_0(X_{all}) = \mathbb{E}[Y|D=0, X_{all}]$.
- It is easy to show that, under unconfoundedness and common support,

$$ATE = \mathbb{E}[m_1(X_{all}) - m_0(X_{all})]$$

- Learning about these m functions is essentially a prediction problem!
- Denote their estimate fitted valued by $\widehat{m}_d(X_{all})$, $d \in \{0,1\}$. Then,

$$\widehat{ATE}_{ra} = \frac{\sum_{i=1}^{n} (\widehat{m}_1(X_{all,i}) - \widehat{m}_0(X_{all,i}))}{n}$$

- Similar to "T-learner" in ML language.

ATE under unconfoundedness: Inverse Probability Weighting

- One can adopt an alternative approach that requires estimating only one unknown function (instead of two)
- Let $p(X_{all}) = \mathbb{E}\left[D|X_{all}\right] = \mathbb{P}(D=1|X_{all})$ denote the propensity score
- $p(X_{all})$ gives the probability a unit with features X_{all} being treated.
- We can get an alternative formulation for the ATE based on $p(X_{all})$:

$$ATE = \frac{\mathbb{E}\left[\frac{D}{p(X_{all})}Y\right]}{\mathbb{E}\left[\frac{D}{p(X_{all})}\right]} - \frac{\mathbb{E}\left[\frac{1-D}{1-p(X_{all})}Y\right]}{\mathbb{E}\left[\frac{1-D}{1-p(X_{all})}\right]}$$

- Learning about p is a classification/prediction problem.
- IPW-based estimates of ATE can be obtained using the "plug-in" approach.

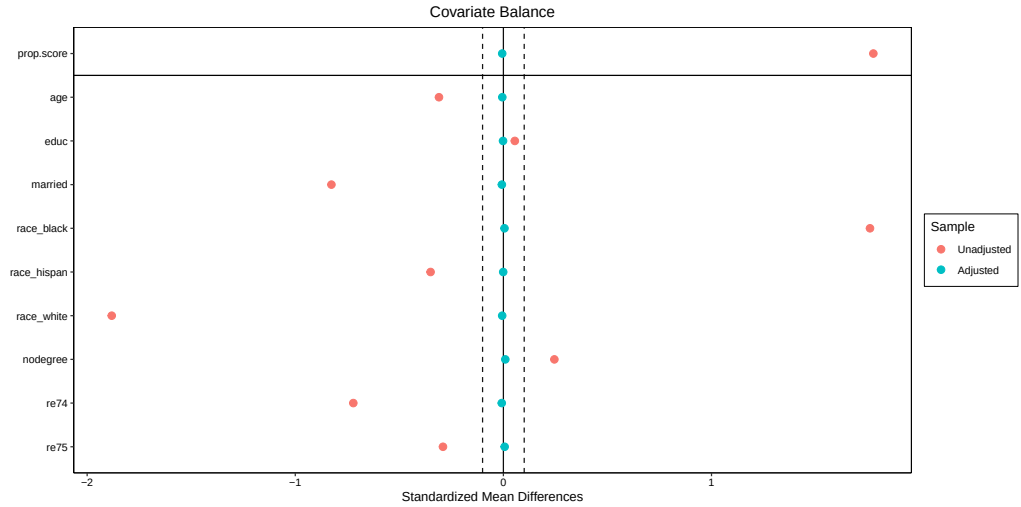
Unconfoundedness: Sanity check about assumptions

- Unconfoundedness allows treated and untreated units to be **different** but only in terms of observables.
- We usually cannot directly test it, but there are some falsification checks that we can do.
- Under unconfoundedness, we have that for any $X \subset X_{all}$

$$\frac{\mathbb{E}\left[\frac{D}{p(X_{all})}X\right]}{\mathbb{E}\left[\frac{D}{p(X_{all})}\right]} = \frac{\mathbb{E}\left[\frac{1-D}{1-p(X_{all})}X\right]}{\mathbb{E}\left[\frac{1-D}{1-p(X_{all})}\right]}$$

- Good practice to check for this
- **But good balance does not mean the design is valid: remember unobservables!**

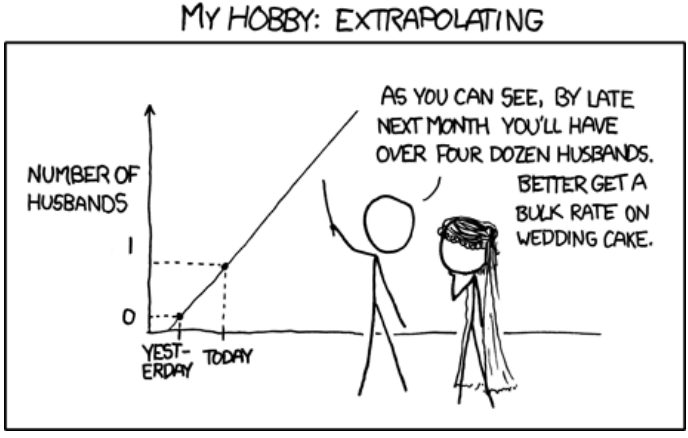
Example of balance check with Lalonde data



Unconfoundedness: Sanity check about assumptions

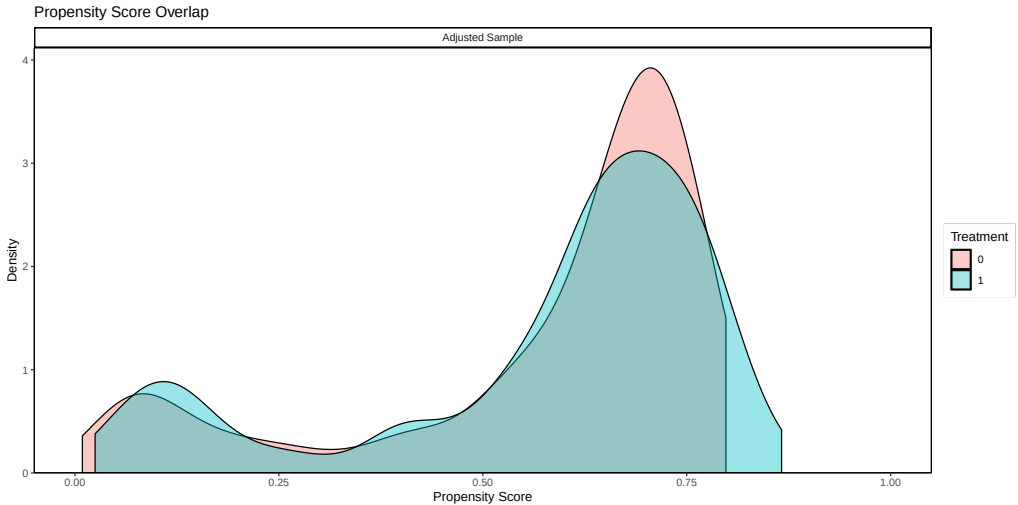
- Another important check is to ensure that we have a good overlap.
- Plot the density of propensity score among treated and untreated units to see what is going on.
- Overlap is important to assess whether we need to rely on “extrapolation” or not.
- Also to assess whether trimming or winsorization is needed (if one does not want to extrapolate).
- Weak overlap also leads to all types of non-standard inference issues.
  - If you are curious, check Ma, Sant’Anna, Sasaki and Ura (2023) for some work-in-progress.

Issues with extrapolation



Source: https://xkcd.com/605/

Example of Overlap Check with Lalonde data



Causal ML under Unconfoundedness

- The previous section suggested two different paths that we could follow to leverage ML for Causal inference
- One based on the regression adjustments procedure to get

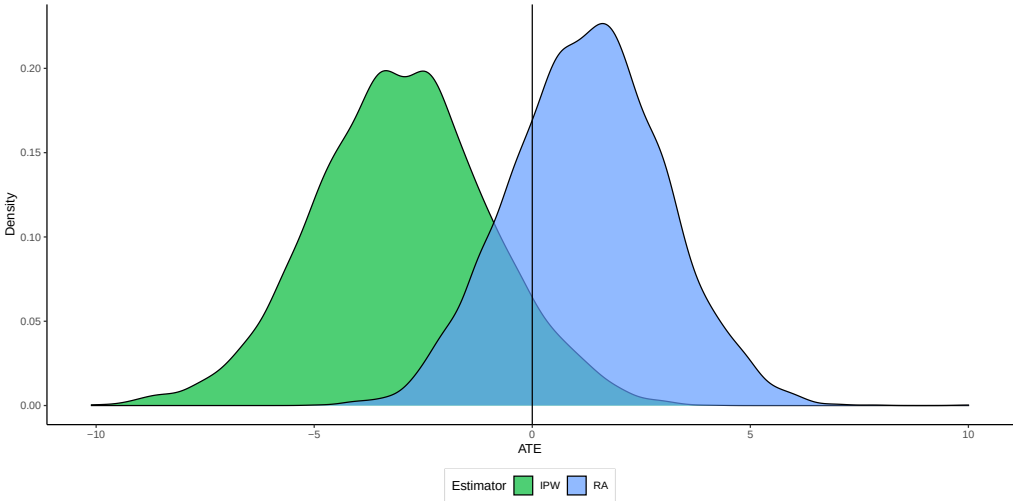
$$\widehat{ATE}_{ra} = \mathbb{E}_n \left[\widehat{m}_1(X_{all,i}) - \widehat{m}_0(X_{all,i}) \right]$$

- The other is based on the IPW procedure to get

$$\widehat{ATE}_{ipw} = \frac{\mathbb{E}_n \left[\frac{D}{p(X_{all})} Y \right]}{\mathbb{E}_n \left[\frac{D}{p(X_{all})} \right]} - \frac{\mathbb{E}_n \left[\frac{1-D}{1-p(X_{all})} Y \right]}{\mathbb{E}_n \left[\frac{1-D}{1-p(X_{all})} \right]}$$

- But does this work?

Causal ML Plug-in does not work: illustration from fake data



ML is not magic, and it does make mistakes

- Although ML is very good for prediction purposes, Causal inference is hard!
- We are using ML (in this case, random forest) to learn about $p(\cdot)$, $m_1(\cdot)$ and $m_0(\cdot)$.
- In reality, ML can approximate these functions well, but it will make "mistakes", at least in finite samples!
- **But what if I tell you that there is a "better" way to use ML?**

Doubly Robust Formulation

- Instead of picking between RA and IPW approach, the idea is to combine both as

$$
\begin{aligned}
ATE \quad = \quad & \left(\mathbb{E}\left[m_1(X_{all})\right] + \frac{\mathbb{E}\left[\frac{D}{p(X_{all})}\left(Y - m_1(X_{all})\right)\right]}{\mathbb{E}\left[\frac{D}{p(X_{all})}\right]} \right) \\
& - \\
& \left(\mathbb{E}\left[m_0(X_{all})\right] + \frac{\mathbb{E}\left[\frac{1-D}{1-p(X_{all})}\left(Y - m_0(X_{all})\right)\right]}{\mathbb{E}\left[\frac{1-D}{1-p(X_{all})}\right]} \right)
\end{aligned}
$$

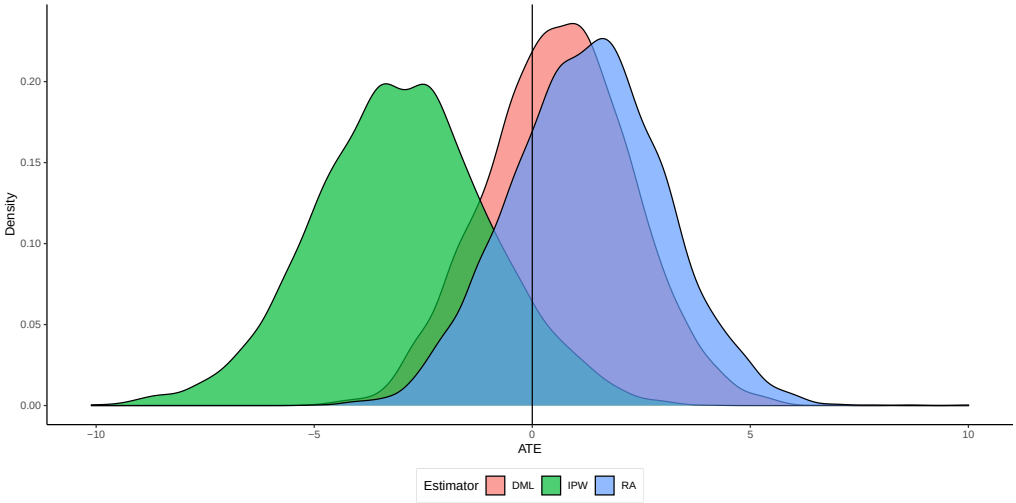
Doubly Robust Estimator aka DML

- Instead of picking between RA and IPW approach, the idea is to combine both as

$$\widehat{ATE} = \left(\mathbb{E}_n\left[\widehat{m}_1(X_{all})\right] + \frac{\mathbb{E}_n\left[\frac{D}{\widehat{p}(X_{all})\left(Y - \widehat{m}_1(X_{all})\right)}\right]}{\mathbb{E}_n\left[\frac{D}{\widehat{p}(X_{all})}\right]} \right)$$

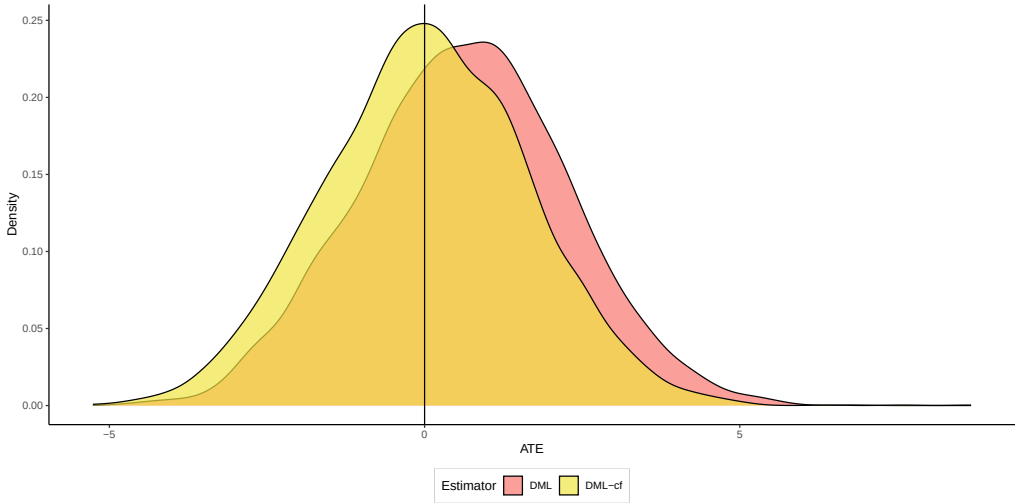
$$- \left(\mathbb{E}_n\left[\widehat{m}_0(X_{all})\right] + \frac{\mathbb{E}_n\left[\frac{1-D}{1-\widehat{p}(X_{all})}\left(Y - \widehat{m}_0(X_{all})\right)\right]}{\mathbb{E}_n\left[\frac{1-D}{1-\widehat{p}(X_{all})}\right]} \right)$$

DML: illustration from simulated data



We can also use cross-fitting to avoid (potential) overfitting

DML with Cross-fitting: illustration from simulated data



Using ML for estimating ATE

- If you want to use ML for Causal Inference, you should use the Doubly Robust formulation.
- More precisely, you need to use estimators that satisfy the Neyman-Orthogonality condition; see, e.g., Belloni, Chernozhukov, Fernández-Val and Hansen (2017), Athey and Wager (2018), Athey et al. (2019).
- You should also favor cross-fitting to avoid regularization bias
- Similar strategies are available to estimate heterogeneous treatment effects, too; see, e.g., Athey et al. (2019).
  - This is what is behind the scenes with many industry causal inference applications!

Causal Inference with Observational Data

- I hope that you are now fully aware of where we are going!
- In the rest of this course, we will discuss how we can do all this, in practice!
- We will mostly talk about experiments and unconfoundedness, but there are many other methods for observational data
  - Difference-in-Differences
  - Instrumental Variables
  - Synthetic Controls
  - Regression Discontinuity
- Many of the ideas we will cover will apply there, too

1. Why do we care?
   Predictions vs. Causality
2. Mapping questions into causal parameters
   - Average causal effect parameters
3. How can I actually learn/estimate these causal parameters?
   - A/B test and RCTs for the rescue
4. What if we do not have an experiment?
   - Unconfoundedness
5. How can we use Machine Learning for causal inference?
6. Closing Remarks

Athey, Susan and Stefan Wager, "Estimation and Inference of Heterogeneous Treatment Effects using Random Forests," *Journal of the American Statistical Association*, 2018, *113* (523), 1228 – 1242.

_ **, Julie Tibshirani, and Stefan Wager**, "Generalized random forests," *The Annals of Statistics*, 2019, *47* (2), 1148 – 1178.

Belloni, Alexandre, Victor Chernozhukov, Iván Fernández-Val, and Christian Hansen, "Program Evaluation and Causal Inference With High-Dimensional Data," *Econometrica*, 2017, *85* (1), 233–298.

Chernozhukov, Victor, Mert Demirer, Esther Duflo, and Iván Fernández-Val, "Generic Machine Learning Inference on Heterogenous Treatment Effects in Randomized Experiments ," *arXiv:1712.04802*, 2022.

Ma, Yukun, Pedro H. C. Sant'Anna, Yuya Sasaki, and Takuya Ura, "Doubly Robust Estimators with Weak Overlap," *arXiv:2304.08974*, 2023.