Causal Inference Crash Course: Foundations

Julian Hsu

Causal Inference Series

1) **Foundations**
2) Defining Some ATE/ATET Causal Models
3) ATE/ATET Inference, Asymptotic Theory, and Bootstrapping
4) Best Practices: Outliers, Class Imbalance, Feature Selection, and Bad Control
5) Heterogeneous Treatment Effect Models and Inference
6) Difference-in-Difference Models for Panel Data
7) Regression Discontinuity Models
8) Arguable Validation

Overview

- This is intended to be the first in a series of causal inference
- This presentation will cover:
  - The fundamental problem of causal inference
  - Necessary assumptions for valid causal inference
  - Working example with simulated data
- The intended audience is a scientist familiar with basic statistics and unfamiliar with causal inference

Fundamental problem of causal inference

- Suppose we want to estimate the return of investment of going to college on career earnings.
- Data:
  - Earnings of college-goers: $Y_i(T = 1)$
  - Earnings of non-college-goers: $Y_i(T = 0)$
- Can we just compare the earnings of college-goers and non-college-goers?
  - Naïve difference: $Y_i(T = 1) - Y_i(T = 0)$

Do we know what happens in the counterfactual situation?

- We cannot, because college-goers tend to be different from non-college-goers.
- In other words, we do not think that the observed earnings of non-college-goers accurately represents the earnings of college-goers if they did not go to college, and vice versa.
  - In other words, correlation is not causation.
- We formalize this with the “potential outcomes” mental model (aka Neyman-Rubin causal model).

Potential outcomes framework

- There is a difference between what we observe, and what we wish we could observe.

	If they did not go to college	If they went to college
Non-College Goers	$Y_i(0,0)$ observed	$Y_i(0,1)$ counterfactual
College Goers	$Y_i(1,0)$ counterfactual	$Y_i(1,1)$ observed

- We only observe $Y_i(0,0)$ and $Y_i(1,1)$ but we want to observe the counterfactuals.
- If we think that $Y_i(1,1) - Y_i(0,0)$ is the real effect of going to college, then we are also assuming that $Y_i(1,1) = Y_i(0,1)$, and $Y_i(0,0) = Y_i(1,0)$.

Selection bias

- Why would $Y_i(1,1) \neq Y_i(0,1)$? If college goers did so because they believed they would get more out of it than those who did not go.
  - College attendees may be higher ability and thus get more out of college
- This is called selection bias
- Causal inference models can address selection bias – under certain conditions

Causal inference and reporting

- The usual metric for reporting purposes is the Average Treatment Effect on the Treated (ATET).
  - ATET answers “What is the impact of going to college for college-goers?”
  - aka: $Y_i(1,1) - Y_i(1,0)$
- Causal inference cares about the Average Treatment Effect (ATE) and ATET.
  - ATE answers “What is the impact of going to college?”
  - aka $\mathrm{E}_1(Y_i(1,1), Y_i(0,1)) - \mathrm{E}_0[Y_i(0,0), Y_i(1,0)]$
  - where E_l is the weighted average of observed and counterfactuals for $T = l$.

Three necessary assumptions for valid causal inference

1. Unconfoundedness assumption[1]: treatment status is random once we condition based on pre-treatment features X_i
   - Conditioning based on X_i is equivalent to comparing synthetic twins
   - AKA – there is no omitted variable bias

- Setup for an Example:
  - Observed outcome: $\mathrm{Y} = \mathrm{f}(\mathrm{X}_1, X_2, \psi) + \tau\, T(X_1, X_2, \epsilon)$
  - $\mathrm{f}(\mathrm{X}_1, X_2, \psi)$ is a function that determines outcome for control accounts
  - τ is the treatment effect
  - $T(X_1, X_2, \epsilon)$ is a that determine whether observation is in treatment or control

[1] aka selection-on-observables, conditional independence, ignorability, conditional independence assumptions.

Scenario 1: the ideal

- Observed outcome: $Y = f(X_1, X_2, \psi) + \tau\, T(X_1, X_2, \epsilon)$
- X_1, X_2 are observed.
- ψ, ϵ are not observed, but are independent of each other.
- Therefore, variation in treatment and the outcome conditional on X_1, X_2 are independent of each other.
  - aka $\mathrm{cov}(Y|X_1, X_2,\ T|X_1, X_2) = 0$.
- Then conditioning on X_1, X_2 yields an unbias estimate of τ.

Scenario 2: not ideal

- Observed outcome: $Y = f(X_1, X_2, \psi) + \tau\, T(X_1, X_2, \epsilon)$
- X_1, X_2 are observed.
- ψ, ϵ are not observed, but are correlated.
- Therefore, variation in treatment and the outcome conditional on X_1, X_2 are not independent of each other.
  - aka $\mathrm{cov}(Y|X_1, X_2,\ \ T|X_1, X_2) \neq 0$.
- Then conditioning on X_1, X_2 yields a bias estimate of τ
  - Bias depends on how correlated ψ, ϵ are.

So what happens in the not ideal scenario?

- You will still get an estimate for τ, but it will be biased.
- Recall this ψ, ϵ are not observed, but are correlated with each other.
- You cannot identify what portion for $\hat{\tau}$ is due to the true treatment effect τ, and what is due to ψ or ϵ.

How do we know if we are in the ideal or not ideal scenario?

- You don't know.
- Whether the unconfounded assumption is met depends on whether the unobserved ψ, ϵ are correlated. But ψ, ϵ are unobserved by definition.
- So what do we do?
- We rely on our knowledge of the causal inference problem so that we can safely assume that ψ, ϵ are not correlated.
- We can also gather suggestive evidence that unconfoundedness is true.

Valid causal inference relies on having the "perfect" but "imperfect" prediction problem

- Observed outcome: $Y = f(X_1, X_2, \psi) + \tau\, T(X_1, X_2, \epsilon)$
- Imagine if you observed $X_1, X_2, \psi,$ and ϵ, so you could perfectly predict Y and T. Can you estimate τ?
- No – because you cannot figure out how much of Y is due to $f()$ and due to $\tau\, T()$.
- In other words, there is no variation in treatment and the outcome conditional on X_1, X_2. So there is no variation to compare to estimate τ.

Randomized control trials / Experiments / AB tests

- When treatment is unconditionally random, then the unconfoundedness assumption is automatically satisfied.

- This means a t-test is an unbiased estimate, but controlling for features X results in a more precise estimate.
  - This is because a t-test ignores variation correlated with X

Three necessary assumptions for valid causal inference - SUTVA

2. Stable unit treatment value assumption: control units do not influence treatment units.
   - A violation of this would be college-goers help non-college-goers get higher paying jobs
- We can only test this if we observe how treatment and control observations are connected. (Network data)
- In practice, this is usually assumed away because of how difficult it is to get network data.

Estimating treatment effects using matching

- We can use the unconfoundedness assumption for causal inference.

- Intuitively: Let's find treatment accounts and control accounts that have the same X_i values. Using the unconfoundedness assumption, conditional on having same X_i values, treatment is random.

Exact and KNN Matching Examples

- For example, suppose we only have one feature $\mathrm{X} \sim U[1,4]$.

- Exact Matching: We would find treatment and control accounts that have the same value of X. You assume that accounts with the same value of X are randomly assigned treatment status.

- KNN Matching: Instead of having the exact value of X, we could match treatment and control accounts that have similar values of X.

Matching leads to the curse of dimensionality

- But what if we have a lot of pre-treatment features X_i?
- For example, suppose we have ten pre-treatment features, each taking a value of 1,2,3,or 4. Then we have 4^{10} = 1,048,576 possible combinations.
- We can't practically find exact, or approximate, matches across a lot of X_i. This is a dimensionality problem.

* Rosenbaum, P. and Rubin, D. 1983. "The central role of the propensity score in observational studies for causal effects." Biometrika. Vol 70-1.

Propensity Score Matching

- Rosenbaum and Rubin (1983) find that instead of matching on each value for each $\mathrm{X_i}$, we can match on one feature instead.

- Match on the predicted probability of treatment, $\hat{P}(X_i)$, known as the propensity score.
  - Propensity score model can be trained from an ML model, as long as you have the treatment label information

- Intuitively, conditional on the propensity score, treatment status is random. This uses the unconfoundedness assumption.

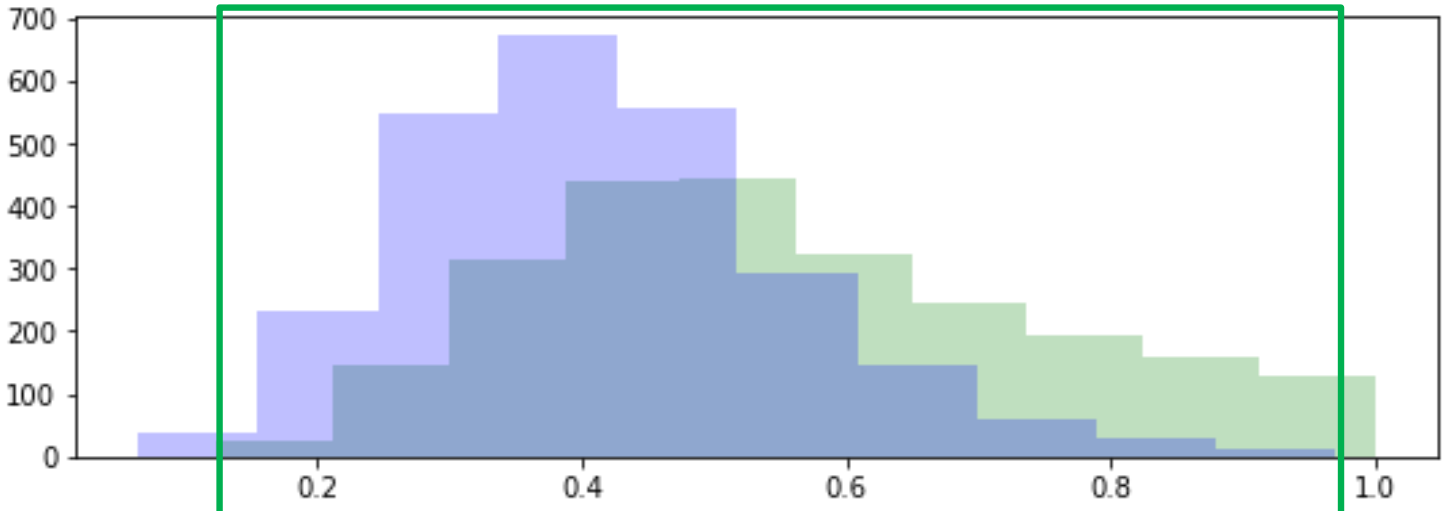
- For example, you find a treatment account with $\hat{P}(X_i) = 0.50$ and a control account with $\hat{P}(X_i) = 0.50$. For these accounts, whether they are in treatment or control is random, given that they have the same $\hat{P}(X_i) = 0.50$.

Three necessary assumptions for valid causal inference - Overlap

3. Overlap assumption: we can find control and treatment accounts with similar propensity score $\hat{P}(X_i)$.

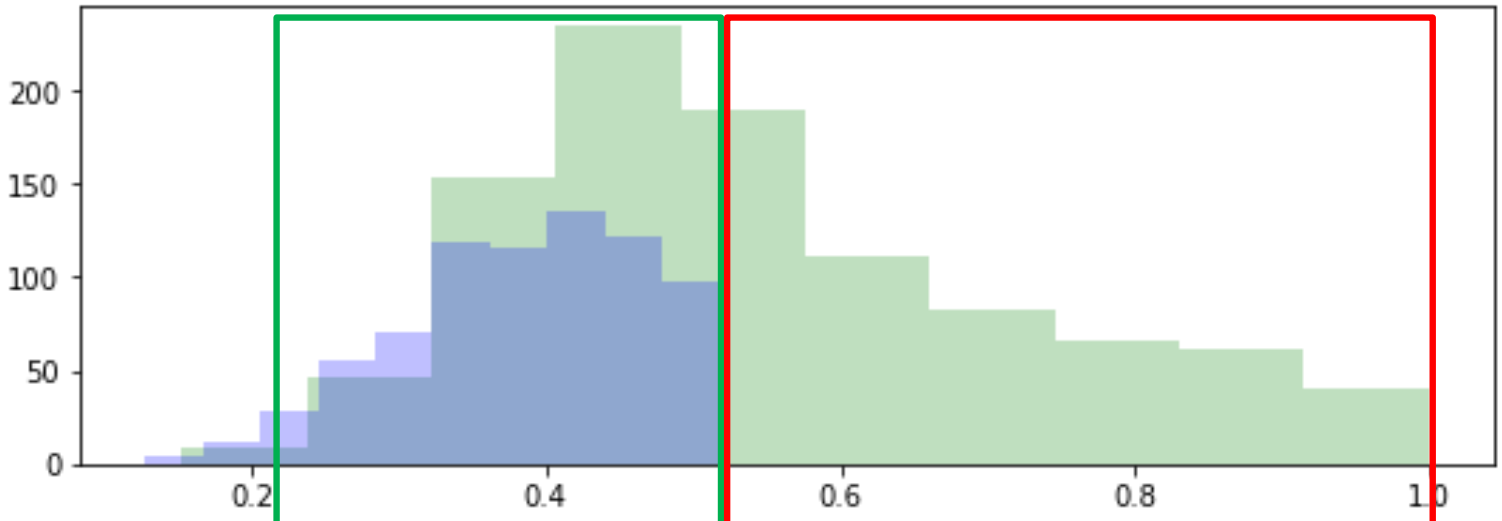
   - This one we can verify if we have the true propensity score. We can plot the distribution of propensity scores for treatment and control groups, and see if there is reasonable overlap.

Ideal scenario



Control and treatment accounts overlap in propensity stores, so there are plenty of control and treatment accounts to compare.

Not ideal scenario



Plenty of control and treatment accounts to compare.

No control accounts with propensity scores here, so no direct comparisons with treatment accounts.
Estimates for these treatment accounts are extrapolations.

Propensity-based models are the most common models

- Many causal models are versions of propensity score matching:
  - Ordinary least squares (OLS)
  - Propensity Binning with Regression adjustment (CBA's DSI 2.0)
  - Inverse propensity weighting
  - Double machine learning (CBA's DSI 3.0)
- They all rely on the same assumptions.

Example with Simulated Data

- Jupyter Notebook link: [link]
- To make sure we can tell whether the average treatment effect estimate is correct, we will use fake data.
- We will setup our fake data, and show the bias under different causal models and whether the assumptions for causal inference are met.

Setup with fake data

- We have four iid features, $\mathrm{x}_1, \mathrm{x}_2, \mathrm{x}_3$, and x_4 from normal distributions.
- The observed outcome is:
  - $Y = f(x_1, x_2, x_3, x_4, \psi) + \tau \times T(x_1, x_2, x_3, x_4, \epsilon)$
  - Where (ψ, ϵ) are independent.

We study three scenarios

1. (Ideal scenario) We observe x_1, x_2, x_3, x_4
2. (Unconfoundedness assumption violated) We only observe x_1, x_2, x_3
3. (Overlap assumption violated) We observe x_1, x_2, x_3, x_4, but control accounts with certain $\hat{p}(x_1, x_2, x_3, x_4)$ values are missing

We study four models

- We estimate ATE with four different causal models:
  1. OLS
  2. Propensity Binning with Regression adjustment
  3. Double machine learning - Partial Linear Model
  4. Double machine learning - Interactive Regression Model
- Let's skip the model definitions for now. Theoretically they should give the same results; but estimates will vary over context.
- In each model, we condition on $\mathrm{x}_1, \mathrm{x}_2, \mathrm{x}_3, \mathrm{x}_4, \mathrm{x}_1^2, \mathrm{x}_2^2, \mathrm{x}_3^2$, and x_4^2 for flexibility. We could condition on even more transformations of $(\mathrm{x}_1, \mathrm{x}_2, \mathrm{x}_3, \mathrm{x}_4)$ for additional flexibility.

Bias in ATE Estimate across five models and three scenarios

- Showing the average bias = $(\tau - \hat{\tau})$ across 500 simulated datasets, each with the same τ value. For reference, $\tau = 5$.

	Ideal Scenario	Unconfoundedness Violated	Overlap Violated
OLS	0.003	0.003	0.722
Propensity Binning w/ Regression Adjustment	-0.062	0.115	-1.722
DML – Partial Linear	0.040	0.248	0.909
DML – Interactive Regression	0.036	0.292	2.186

Conclusion

- This presentation covered some fundamentals for causal inference:
  - The fundamental causal inference problem
  - Assumptions for unbiased estimates
  - Working example

- The following slide has a list of future topics. Please let me know what else interests you!

Appendix Slides

Details on Modeling Approaches

- The DML-IRM and IPW approaches exclude observations with estimated propensity scores below 0.001 and 0.999 to avoid dividing by a small number, which would inflate estimates.
- Except for OLS, all nuisance functions to predict the outcome and treatment status are estimated using four-fold cross-validated LASSO and Logistic regressions.