Causal Inference Mid-Term Review

Fall 2018

Intro to Potential Outcomes

- Defining States
- **SUTVA**
- **Potential Outcomes**
- **ATE Definition**
- **CATEs**
- Biases
  - Baseline
  - Differential Treatment Effect
- Ignorability

Defining States

- “Fine articulation” -- precisely defining the variables in the graph
- Benefits:
  - Reproducible results
  - Possibly fewer dependencies with other variables

Potential Outcomes

- $Y_i(d)$ is the outcome unit i would have in the treatment state d, the “potential outcome”
- We assume each unit has some specific value it would take on in each state
- We can only ever measure one potential outcome for each unit: the outcome it actually takes (due to its treatment state)

SUTVA: Stable Unit Treatment Value Assumption

- The potential outcomes for each unit should be independent of the treatment assignments of the other units:
- $Y_i(1), Y_i(0) \perp D_j$, where the $Y_i(d)$ are the potential outcomes for unit i, and D_j is the assignment for unit j.

Examples:

- Effect of vaccination (D=1) vs. no vaccination (D=0) on whether someone gets a disease (Y=1) or not (Y=0)
- For fixed $D_i = d_i$, if more people are vaccinated, unit i is less likely to get the disease. $Y_i(d_i)$ depends on D_j!
- The result we get from any experiment depends on the size of the treatment group!

ATE: Average Treatment Effect

- The average effect of changing from the control state to the treatment state
- $\boldsymbol{\delta} = E[Y(1) - Y(0)]$, the average difference in potential outcomes
- If $Y_i(d)$ are independent of D, then

  $\boldsymbol{\delta} = E[Y|D=1] - E[Y|D=0] \overset{\text{def}}{=} \boldsymbol{\delta}_{\text{NAIVE}}$

- Generally, $\boldsymbol{\delta} \neq \boldsymbol{\delta}_{\text{NAIVE}}$ due to confounding, which causes $Y(d) \pm D$

CATEs: Conditional Average Treatment Effects

- We can ask what the average treatment effect is for subsets of the population
- Those subsets can be created by stratifying on anything!
- ATE conditional on X=x is E[Y(1) - Y(0)|X=x].
- A common choice is X = D, which we call (for binary D) the ATT (average treatment effect on the treated) for D=1, and ATC (average treatment effect on the control) for D=0.
- ATT is the treatment effect for people who normally take the treatment (remember, we’re thinking of observational data).

Ignorability

- When $Y(1), Y(0) \perp D$, we call treatment assignment “ignorable”. That is, you can ignore the treatment assignment mechanism (the usual source of bias, looking only at one side of the confounding fork), and use the naive estimator.
- If $Y(1), Y(0) \perp D \mid Z$, conditional on some set Z, we say treatment assignment is “conditionally ignorable”. That is, Z blocks the back-door paths between Y and D.

Intro to Causal Graphs

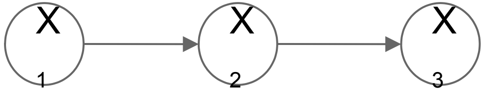
- Graphs to represent joint distributions (purely statistical)
- Graphs to represent causal systems (causal and statistical)
- **Joint Distribution Factorization**
- **The do(X=x) operation**
- **Statistical dependence from paths (intuition)**
  - **Forks**
  - **Colliders**
  - **Chains**
- **Blocking and d-separation**
- **The back-door criterion**

Factorizing Graphs

- Given a DAG, G, we can factorize a joint distribution over G like

  $P(X_1, \dots, X_N) = \prod_{i=1}^{N} P(X_i|Par_G(X_i))$

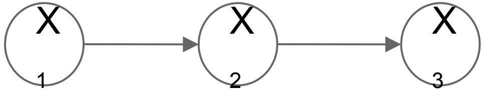
- Comes from applying conditional independence assumptions (consistent with the graph) to the chain rule for conditional probability.



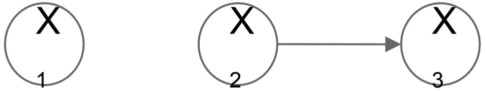
$P(X_1, X_2, X_3) = P(X_3|X_2)P(X_2|X_1)P(X_1)$

The do(X_i=x_i) operation

- The graph represents the system that produces observational data.
- We'd like a graph that represents experimental data.
- We can cut the causes of X_i and fix its value to X_i=x_i to describe controlling X_i to set its value, i.e. in an experiment
- We can modify the factorization to get a new joint for the experimental data!
  - Set the $P(X_i|Par(X_i))$ to 1 when you intervene on X_i
  - Set all terms conditional on X_i to the value x_i



$$P(X_1, X_2, X_3) = P(X_3|X_2)P(X_2|X_1)P(X_1)$$



$$P(X_1, X_3|do(X_2=x_2)) = P(X_3|X_2=x_2)\ 1\ P(X_1)$$

$$= P(X_1, X_2=x_2, X_3) / P(X_2| X_1)$$

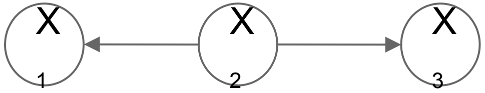
So $P(X_2, X_1) > 0$ is a requirement!!!

Statistical Dependence from Paths

- Statistical dependence is a result of causal structure!
- Three main structures: forks, colliders, and chains.

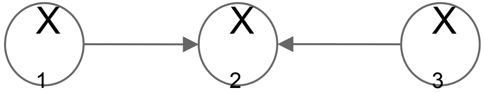
Statistical Dependence from Paths: Forks

- Forks: X_2 a common cause of X_1 and X_3
- Like hot temperature driving both lemonade sales and crime: lemonade sales and crime will be correlated, even though they don't cause each other!
- On days with the same temperature, variation in lemonade sales and crime rates are independent of each other!
- Lemonade sales and crime rates are conditionally independent, given the middle node of the fork.
- X_1 and X_3 are statistically dependent, but not causally dependent.



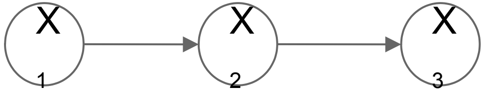
Statistical Dependence from Paths: Colliders

- X_1 and X_3 both cause X_2
- X_1 and X_3 are *statistically independent,* (and not causally dependent either.)
- It's like social skills and math skills driving school admission
- If we know someone is bad at math, and they're a student at the school, then they must have good social skills.
- Conditional on being a student, math skills and social skills are negatively correlated
- While independent in the general population, X_1 and X_3 are *conditionally dependent!* Conditioning on the collider results in statistical dependence!



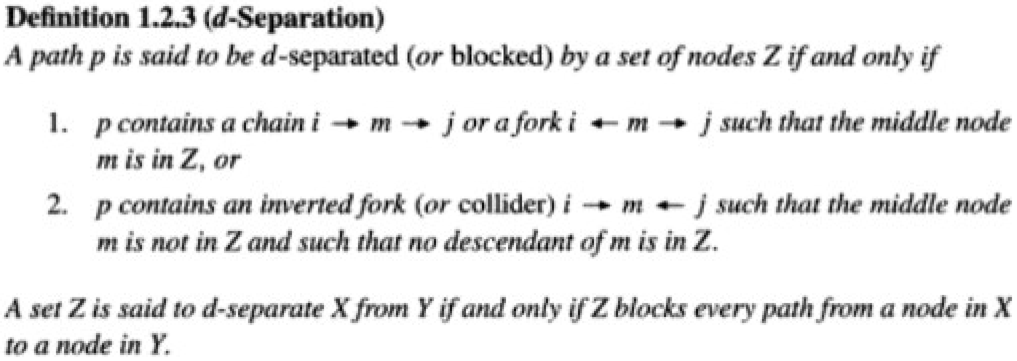
Statistical Dependence from Paths: Chains

- X_1 causes X_2 which causes X_3. X_1 only causes X_3 through X_2
- This is like power outages (X_1) causing the lights to turn off (X_2), causing darkness.
- X_1 and X_3 are causally dependent, and statistically dependent.
- Once we know the lights are off, we know all that we need to know to say whether it’s dark. The power outage doesn’t provide any new information.
- $P(X_3|X_2) = P(X_3|X_2, X_1)$, or equivalently, $P(X_3,X_1|X_2) = P(X_3|X_2)P(X_3|X_1)$, or X_1 and X_3 are conditionally independent, given X_2!



Blocking and d-separation

- A path is “blocked” by a set of variables Z when the nodes on either end of it are statistically independent conditional on Z.
- Let’s do some examples! (on the board)



Definition 1.2.3 (*d*-Separation)

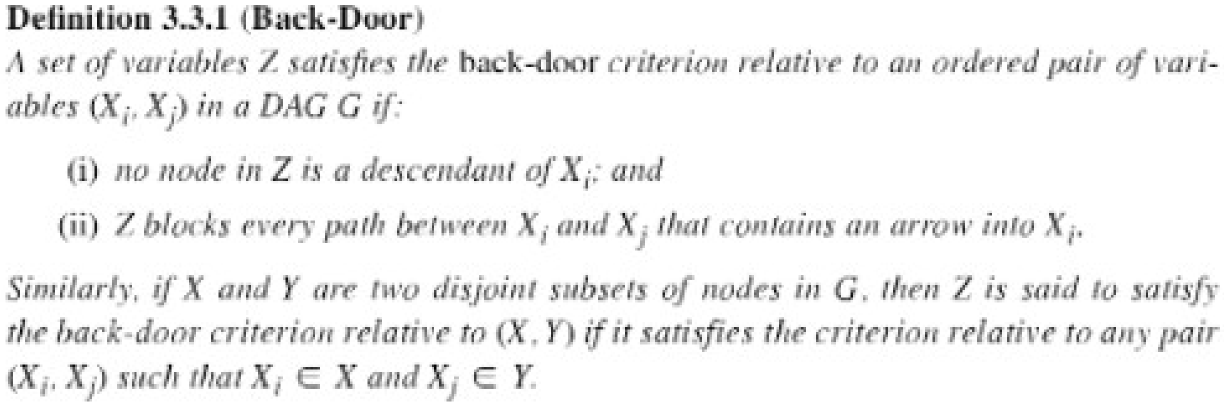
*A path p is said to be d-*separated (*or* blocked) *by a set of nodes Z if and only if*

1. *p contains a chain i → m → j or a fork i ← m → j such that the middle node m is in Z, or*
2. *p contains an inverted fork* (*or* collider) *i → m ← j such that the middle node m is not in Z and such that no descendant of m is in Z.*

A set Z is said to d-separate X from Y if and only if Z blocks every path from a node in X to a node in Y.

The back-door criterion

- We block non-causal paths
- We don’t block causal paths
- We don’t unblock non-causal paths.
- All that is left is causal dependence!



Definition 3.3.1 (Back-Door)

A set of variables Z satisfies the back-door *criterion relative to an ordered pair of variables* (X_i, X_j) *in a DAG G if:*

(i) *no node in Z is a descendant of* X_i; *and*

(ii) *Z blocks every path between* X_i *and* X_j *that contains an arrow into* X_i.

Similarly, if X and Y are two disjoint subsets of nodes in G, then Z is said to satisfy the back-door criterion relative to (X, Y) if it satisfies the criterion relative to any pair (X_i, X_j) *such that* $X_i \in X$ *and* $X_j \in Y$.

Matching

- Idea of matching
- ATT and ATC
- Pros/Cons of matching methods
  - Exact Matching
  - K Nearest Neighbors
  - Caliper (or distance-based) matching
  - Weighting
- Identifying CATEs when ATE isn’t identifiable
- Checking the support
  - Parametric models vs. Exact matching
- Conditional Ignorability
- Propensity Scores and blocking

Basic Idea

- We need to condition on Z to block back door paths
- If we can find treatment and control units with the same Z, we can find a noisy estimate of δ
- If we average them together, we can get a good estimate!
- All we need is a statistic of Z: the propensity score

ATT and ATC

- We covered these in more depth here, because we matched control units to treated ones (to get the ATT) and vice versa (to get the ATC).
- ATT is the treatment effect for people who usually take the treatment
- ATC is the treatment effect for people who don’t usually take the treatment.
- (remember: we’re thinking of *observational* rather than *experimental* data)

Pros and Cons of Matching Methods

All methods:

- Cons:
  - We have to find a set Z to control for!
  - We have to have measured all the Z
  - $P(D,Z) > 0$ is still a requirement
    - we have to have units in the treatment and control state take on all values of Z that the other group takes on
    - slightly more flexible with parametric methods
  - Error bars can be hard to calculate
  - You often won’t use all of the data.
- Pros:
  - We can use back door sets when assignment is conditionally ignorable to get ATE estimates!
  - You can use doubly robust estimation to more

Matching Pros and Cons: Exact Matching

- Pros
  - Non-parametric: You don’t have to choose a model (which might be wrong)
- Cons
  - More strict requirement that P(Z,D) > 0, but you can still estimate ATC and ATT in some cases when it isn’t.
  - The more variables you try to control for, the less likely P(Z,D) > 0 (curse of dimensionality)

Matching Pros and Cons: Nearest Neighbors

Pros:

- Matching on the propensity score, so the data density issues aren’t as bad!
  - As long as you’re okay extrapolating the model where P(D,Z) = 0, you’re okay.

Cons:

- You can have high variance when the number of neighbors is small, but higher bias when it’s big
  - there are generally many parameters to adjust to trade bias and variance.
- Model misspecification leads to bias.

Matching Pros and Cons: Caliper Matching

Pros:

- Matching on the propensity score, so the data density issues aren't as bad!
    - As long as you're okay extrapolating the model where P(D,Z) = 0, you're okay.

Cons:

- There can be high variance if there are few neighbors within the caliper, but higher bias if the caliper is larger.
    - You have to adjust the paramters
- Model misspecification leads to bias.

Matching Pros and Cons: Weighting

Pros:

- Matching on the propensity score, so the data density issues aren’t as bad!
  - As long as you’re okay extrapolating the model where P(D,Z) = 0, you’re okay.
- Fewer parameters for the actual matching process (no caliper or number of neighbors)

Cons:

- Model mis-specification leads to bias.
- Some data points can be disproportionately impactful.

Regression (part 1)

- Structural functions (y = f_y(Par(Y)))
- **Noise terms and omitted variable bias**
- Types of correlation between the error term (on Y) and assignment (D).
- **Regression with the g-formula**
- Doubly robust estimation

Noise Terms and OVB

Correct regression specification is

$Y = \delta D + \beta Z + \varepsilon$

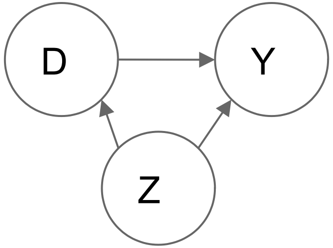
But instead we use

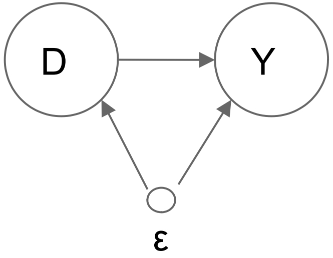
$Y = \delta D + \varepsilon$

Problem: epsilon is correlated with D! We get *omitted variable bias*

(there is confounding, so the naive estimator fails. Take covariances with D to convince yourself.)

Conditioning on a back door set Z renders epsilon independent of D!





Regression with the g-formula

- OLS regression is tricky, because it’s hard to get the model specification right when the treatment effect is heterogeneous.
- Machine learning helps! Just use the g-formula and MSE loss.
- δ = E[Y|do(D=1)] - E[Y|do(D=0)]
- E[Y|do(D=d)] = Σ_Z E[Y|Z=z, D=d] P(Z)
- The first term in the sum can be estimated with a ML model. The sum itself (and the P(Z)) term can be turned into a sum over data.
- E[Y|do(D=d)] = $\Sigma_{i=1}^{N}$ E[Y|Z=z_i, D=d] / N

Regression (part 2)

- Inverse propensity weighting
- Practical issues with controlling-based estimators

Inverse Propensity Weighting

- It turns out even a naive regression can work it we weight with inverse propensity scores!
- The result followed from the g-formula
- We have all the advantages of propensity score estimation methods
  - Extrapolate outside the support of the data
  - Use all of the data for better statistical efficiency
- We also have the advantages of g-formula methods
  - Use more general machine learning models to be less susceptible to mis-specification
  - We can generalize to continuous causal states
- And we can still be doubly robust!
- We can even incorporate survey weights!!!

Examples; Selection bias

- Drawing graphs for systems
- Finding Z for real systems
- Intuition for self-selection bias

Examples

- We covered lots of examples, and used some intuition driven by the BDC to focus our thinking
- We need to block paths with an arrow into the cause, so try to think of all the causes of the causal state, D.
- Assess whether those causes can also (directly or indirectly) cause Y.
- If so, we need to measure and control for them.

Exam Format

- N choose k sections
  - Word problems
  - Some of the harder problems/sections
- Content (likely including, but not limited to...)
  - Questions about simulating data
  - Questions about implementing estimators
  - D-separation problems
  - Back-door conditioning problems
  - Word problems about graph structure
  - Questions about intuition behind the concepts
    - ATT, ATC, and observational vs. experimental data
    - Forks, colliders, and chains; causal vs. statistical dependence.