Causal Inference Talk

roko

October 2022

1 QUESTIONS TO ADDRESS

2 Opening Remarks

1. The fabled "Correlation does not imply causation." statement is pretty well understood.
2. We know of the shark attacks and ice cream sales example.
3. Is it at all possible to identify the causal effect X has on Y? Chapter 19-22 in Advanced Data Analysis covers exactly this.
4. Causal inference actually refers to 2 research questions:
   (a) What is the causal effect of X on Y?
   (b) What is the causal structure existent in an observable dataset?

2.1 Important Definitions

- directed path: a path that can be taken by following the directions of the arrows (aka not backward)
- collider: a node that has 2 arrows pointing into it
- back-door path: A path that connects X to Y that is not direct.
- confounder: a variable that prevents a conditional probability from being causal
- d-separation: a variable S that implies $X \perp\!\!\!\perp Y|S$ on a causal graph
- exogenous: a variable whose measure is determined outside the model and imposed on the model

2.2 Causal Graphical Models

Causal graphical models are DAGs that describe the flows of causality.

Interesting to note that cycles are not allowed in this analysis. This is likely do to the fact that backdoor and front door criterion are impossible to distinguish.

Assumptions of a Causal Graphical Model:

1. There exists a DAG G representing the causal relations among variables.
2. The joint distribution respects the Markov property established in G.
3. Faithfulness. The only conditional independence relationships that exist are the ones indicated by the DAG.

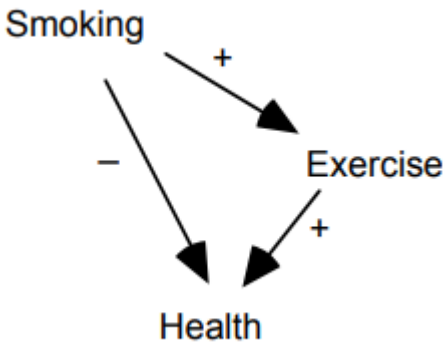

Figure 1: Example of an unfaithful DAG. Suppose that smoking and health are modelled jointly as a multivariate Gaussian. If they are uncorrelated (because the Smoking → Exercise → Health path offsets the direct path), then they are independent. This is unfaithful because there is an independence not directly concluded from the graph.

If the faithfulness assumption is not met, then we do not have an informed way to block off non-causal flows of information.

3 Formalizing Causal Inference

3.1 The do(.) Operator

- The do() operator is a way to perform causal conditioning.
- The do() operator separates a probabilistic prediction from identifying a cause-and-effect relationship.

- The do() notation does not lend itself automatically to a probabilistic calculation. We will outline the conditions needed for this to happen going forward.

3.2 do(.) Sampling vs. Sub-Sampling

Suppose we partition our data into 3 groups: causes (X_c), effects (X_e), and other (X_N).

An important thing to recognize is that simply solving for $\Pr(X_e|X_c = x_c)$ is not enough for causal inference. This is because there exist several ways that the knowledge that $X_c = x_c$ can seep its way into a child node through various back channels. This is the concept of sub-sampling.

The do(.) changes the structure of the graph before a probability is "calculated". "Calculated" is in quotation marks because it only takes on an analytical form if specific conditions are met. We can calculate exogenous

The do(.) operator changes the graph like so:

1. Eliminate any arrows coming in to nodes X_c. (NOTE: this is likely going to include several variables).
2. Fix their values to X_c. (NOTE: this is likely going to include several variables).
3. Calculate the resulting distribution of X_e in the new graph.

If changing X_c from x_c to $x_c^{'}$ changes the distribution of X_e, we say that X_c has a *causal* effect on X_e. This is because we know that is *exclusively* from manipulating X_c that we observed a change in X_e.

The do(.) operator attempts to "exogenize" the variable X_c to analyze its causal effect on X_e.

There are instances where $\Pr(X_e|X_c = x_c) = \Pr(X_e|\text{do}(X_c = x_c))$. When will this happen? When there are no backchannels for X_c to influence X_e in the original graph.

I hope I have established that the main difference between $\Pr(X_e|X_c = x_c)$ and $\Pr(X_e|\text{do}(X_c = x_c))$.

3.3 Identifiability

- To actually calculate $\Pr(X_e|\text{do}(X_c = x_c))$, we'd like to create an experiment with proper data collection. However, this may not feasible either because we cannot control over all the variables we need to or the data collection procedure that would allow us to calculate it via $\Pr(X_e|X_c = x_c)$ is economically/morally unfeasible.
- So, we turn to *identification strategies* to identify causal effects.
- A parameter is said to be identifiable if we can determine its value with infinite data.

- We can identify causal relationships from observational data, but it requires us to *know the causal graph.*
- Example 1: $N(\mu, \sigma)$ is identifiable because it can ONLY happen when μ and σ take on specific values. $N(0, 1)$ observed across 2 distributions means both distributions have the same (μ, σ).
- Example 2: An unindentifiable example would be $N(s_1 + s_2, 1)$. If s_1 and s_2 are parameters of interest you would not be able to identify whether you have (s_1, s_2) = (5,4) or (s_1, s_2) = (6,3) since they both result in the same distribution $N(9, 1)$.

The effect of X on Y is confounded when $\mathbf{Pr}(X_e|\mathbf{do}(X_c = x_c)) \neq \mathbf{Pr}(X_e|X_c = x_c)$.

Goal: We want to have adequate control variables that will block paths connecting X and Y other than the ones which would still exist in the surgically altered graph created by the do() operation.

With adequate control,

$$\Pr(Y|\text{do}(X = x)) = \sum_t \Pr(Y|X = x, \text{Pa}(X) = t)\Pr(\text{Pa}(X) = t)$$

3.4 Adjustments

- These adjustments are important becuase they take our conceptual $P(Y|do(X = x))$ and provide a way to calculate it *exclusively from conditional distributions of our observable data.*
- S, M, and I across all the criteria must be observable (measurable feels more appropriate but it should not be confused with the definition from measure theory) in order for us as statisticians to reasonably condition on them.

3.4.1 Back-Door Criterion

We want to estimate the effect of X on Y. A set of conditioning variables (controls) S satisfies the back-door criterion when...

I) S blocksevery back-door path between X and Y.

II) No node in S is a descendant of X.

If the back-door criterion is satisfied then,

$$\Pr(Y \mid do(X = x)) = \sum_s \Pr(Y \mid X = x, S = s) \Pr(S = s)$$

To prove this, it may be helpful to translate our conditions to mathematical statements.

1. Condition 1 implies $Y \perp\!\!\!\perp \text{Pa}(X)|X, S$.

2. Condition 2 implies $X \perp\!\!\!\perp S|\text{Pa}(X)$

The Entner conditions provide sufficient (but not all necessary) conditions for the back-door criterion to be satisfied. I will skip over these for now, but the advanced reader can refer to pages 440-441.

3.4.2 Front-Door Criterion & Mediators

We want to estimate the effect of X on Y. A set of conditioning variables (controls) M satisfies the front-door criterion when...

I) M blocks all directed paths from X to Y

II) There are no unblocked (aka they are all blocked) back-door paths from X to M.

III) X blocks all back-door paths from M to Y.

If the front-door criterion is satisfied,

$$\Pr(Y \mid do(X = x)) =$$
$$\sum_m \Pr(M = m \mid X = x) \sum_{x'} \Pr\left(Y \mid X = x', M = m\right) \Pr\left(X = x'\right)$$

Again, the proof for this claim is aided by translating our conditions into mathematical statements.

Once an isolated set of mechanisms is found, we know all the ways X actually affects Y and the front door criterion can be utilized.

This can allow us to use sub-mechanisms which *are* isolated when the original mechanism is not.

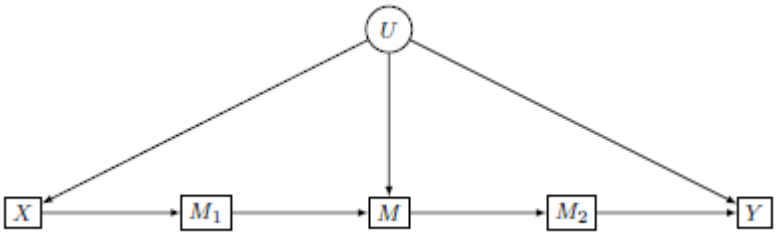

Figure 2: We can identify the causal effect of X on Y in this example. If the original system were only (X, Y, M, and U), then we would not have an identifiable causal effect. However, with M_1 and M_2, we can calculate $\Pr(Y|\text{do}(X = x) = \sum_m \Pr(Y \mid do(M = m)) \Pr(M = m \mid do(X = x)) = \sum_m \left(\sum_{m_2} \Pr(Y = y|\text{do}(M_2 = m_2))\Pr(M_2 = m_2|\text{do}(M = m))\right) * \left(\sum_{m_1} \Pr(M = m|\text{do}(M_1 = m_1))\Pr(M_1 = m_1|\text{do}(X = x))\right)$

3.4.3 Instrumental Variables

instrumental variable: A variable I is an instrument if it influences Y but only through first influencing X which then in turn influences Y.

We want to estimate the effect of X on Y. A variable I is an instrument when there is a set of controls S such that...

I) $I \not\perp\!\!\!\perp X|S$

II) Every unblocked path from I to Y has an arrow pointing into X. ($I \perp\!\!\!\perp Y|S, do(X)$)

I can be important predictor for X, so performing instrumental regression is a combination of two regressions:

1. $X \sim I$
2. $Y \sim \hat{X}$ (I is *deliberately* not included)

The coefficient of Y on $\hat{X}$ is a consistent estimator of the causal effect X has on Y.

If I is a valid instrument, then

$$\Pr(Y \mid do(I = i)) = \sum_x \Pr(Y \mid do(X = x)) \Pr(X = x \mid do(I = i))$$

Instrumental variables can be useful because under linearity assumptions we can get the coefficient of X on $Y \sim X$ when accounting for the confounding variables and (potentially) claim that the effect is causal.

An example of the instrumental linear regression and the issue of confounding is well outlined on page 445.

3.5 Un- and Partial Identifiability

- Surprisingly, when identification is not possible (aka we can't deconfound), it may still be possible bound causal effects. This is called *partial identification*.
- If the list of possible confounders is incredibly large (or even incredibly vague to measure), the identification becomes impossible and all of this is rendered moot.
- To quote the authors, "Often, nothing can be made to make it work." However, we have established when we can and a robust way to think about when a causal statement is an appropriate one to make.

4 Estimating Causal Inference

In the previous section we determined various conditions that are sufficient to identify when $\Pr(Y = y|\text{do}(X = x))$ is calculated using obtainable conditional probabilities.

The point of identification strategies is to reduce the (conceptual) problem of causal inference to that of ordinary statistical inference.

While the adjustment strategies allow us to work with ordinary statistical inference, these probability distributions are not just given to us. We still need to model for some conditional events.

4.1 Effects and Their Calcuations

Having estabilised that under the back-door criterion,

$$\Pr(Y \mid do(X = x)) = \sum_s \Pr(Y \mid X = x, S = s) \Pr(S = s)$$

Then if $\hat{\Pr}(S = s)$ is a consistent estimator of $\Pr(S = s)$ and $\hat{\Pr}(Y = y|X = x, S = s)$ is a consistent estimator of $\Pr(Y = y|X = x, S = s)$, we have that both estimators converge in probability (and therefore in distribution) to their respective estimands. With this information, by Slutsky's theorem,

$$\sum_s \hat{\Pr}(S = s) * \hat{\Pr}(Y = y|X = x, S = s) \xrightarrow{d} \sum_s \Pr(S = s) * \Pr(Y = y|X = x, S = s) = \Pr(Y = y|\text{do}(X = x)$$

The remainder of this subsection will dissect the ways we can obtain a consistent estimator.

(All of the subsequent analysis is still valid if we have the front door criterion because according to the author, "estimating with the front-door criterion amount to doing 2 rounds of back-door adjustment").

4.1.1 Average Effect

A good way to describe the effect X had on Y is throught the average effect, which is defined by $E(Y|\text{do}(X = x)) = \sum_y y\text{Pr}(Y = y|\text{do}(X = x))$.

Remembering our backdoor criterion,

$$\begin{aligned} E[Y \mid do(X = x)] &= \sum_y y \Pr(Y = y \mid do(X = x)) \\ &= \sum_y y \sum_s \Pr(Y = y \mid X = x, S = s) \Pr(S = s) \\ &= \sum_s \Pr(S = s) \sum_y y \Pr(Y = y \mid X = x, S = s) \\ &= \sum_s \Pr(S = s) E[Y \mid X = x, S = s] \end{aligned}$$

4.1.2 Only Estimate the Good Stuff

As previously seen, estimate $\hat{\Pr}(S = s)$ can be a challenging problem. Challenges can include but are not limited to the curse of dimensionality and inability to sample from S.

However, we can utilize the law of large numbers to completely sidestep estimating $\hat{\Pr}(S = s)$. If we have iid samples from S, then

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^{n} \Pr\left(Y = y \mid X = x, S = s_i\right) &\to E(\Pr\left(Y = y \mid X = x, S\right)) \\ &= \sum_s \Pr\left(Y = y \mid X = x, S = s\right) \Pr\left(S = s\right) \\ &= \Pr\left(Y = y|\text{do}(X = x)\right) \end{aligned}$$

So, all we really need a consistent estimator for is $\Pr\left(Y|X, S\right)$.

4.1.3 Average Treatment Effect

In the event that our causal variable X is binary, we can actually identify something called the *average treatment effect*. We can identify one label as a treatment and the other as a control and determine if the treatment has an effect on the outcome variable Y.

$$ATE := E[Y|\text{do}(X = 1)] - E[Y|\text{do}(X = 0)]$$

4.1.4 Calculating ATE

To calculate the ATE, we need to be in possession of a backdoor set of control variables S. This takes us from a conditional with a dooperator in it to a standard condotional probability.

$$
\begin{aligned}
ATE &= \sum_s \Pr(S=s) E[Y \mid X=1, S=s] - \sum_s \Pr(S=s) E[Y \mid X=0, S=s] \\
&= \sum_s \Pr(S=s)(E[Y \mid X=1, S=s] - E[Y \mid X=0, S=s]) \\
&:= \sum_s \Pr(S=s)(\mu(1,s) - \mu(0,s)) = E[\mu(1,s) - \mu(0,s)]
\end{aligned}
$$

Given a sample, we could get an estimator using the sample mean of this function, which would be consistent by the Law of Large Numbers.

$$
ATE \approx \frac{1}{n} \sum_{i=1}^{n} \mu\left(1, s_i\right) - \mu\left(0, s_i\right)
$$

If we could observe $\mu(1, s_i)$ or $\mu(0, s_i)$, we could be done. Unfortunately, the best you can *actually* observe is $Y_i = \mu(x_i, s_i) + \epsilon_i$

Consistent ATE estimators can come from...

1. Regression
2. Grouping
3. Matching

Grouping: The idea is we collect the expected value of the differences of means between two groups. The two groups are $X = 0$ and $X = 1$ and we identify the difference of means for every value $S = s$. If we denote $\mathcal{T}_s$ the set of observations s.t. $X = 1, S = s$ and $\mathcal{C}_s$ the set of observations s.t. $X = 0, S = s$.

$$
\begin{aligned}
&\sum_s \left(\frac{1}{|\mathcal{T}_s|} \sum_{i \in \mathcal{T}_x} Y_i - \frac{1}{|\mathcal{C}_s|} \sum_{j \in \mathcal{C}_x} Y_j \right) \Pr(S=s) \\
&= \sum_s \left(\frac{1}{|\mathcal{T}_s|} \sum_{i \in \mathcal{T}_x} \mu(1,s) + \epsilon_i - \frac{1}{|\mathcal{C}_s|} \sum_{j \in \mathcal{C}_x} \mu(0,s) + \epsilon_j \right) \Pr(S=s) \\
&= \sum_s (\mu(1,s) - \mu(0,s)) \Pr(S=s) + \sum_s \left(\frac{1}{|\mathcal{T}_s|} \sum_{i \in \mathcal{T}_x} \epsilon_i - \frac{1}{|\mathcal{C}_s|} \sum_{j \in \mathcal{C}_x} \epsilon_j \right) \Pr(S=s)
\end{aligned}
$$

The first sum is by definition the expected value of the ATE. The right term is just the sum of weighted noise measurement that are all centered at 0 and will tend to 0 if we take the limit as $n \to \infty$

Matching: The idea behind matching is that we can match values between these groups one-to-one. If we can create n pairs s.t. Y_i is an observation with $X = 1, S = s$ and Y_{i*} is an observation with $X = 0, S = s$, then

$$\frac{1}{n}\sum_{i=1}^{n} Y_i - Y_{i^*} = \frac{1}{n}\sum_{i=1}^{n}(\mu(1,s_i)+\epsilon_i) + (\mu(0,s_{i*})+\epsilon_{i*})$$
$$= \frac{1}{n}\sum_{i=1}^{n}\mu\left(1,s_i\right) - \mu\left(0,s_i\right) + \frac{1}{n}\sum_{i=1}^{n}\epsilon_i$$

The noise sum will go to 0 as $n \to \infty$, and the left summand tends to $E(\mu(1,s)-\mu(0,s))$, which is the ATE.

Grouping and matching have the added benefit of being more accurate locally, but at the expense of boiling down to nearest-neighbor regression, which has complications with growing sample size (n) and matching size (k) (since we can create k matches for each observation).

4.1.5 Propensity Scores

Sufficient statistics can play a helpful role in calculating a causal effect because we can sum over a sufficient statistic rather than a potentially high dimensional S. A sufficient statistic R is defined as a value that provides the same information as S when performing inference for a parameter.

For example, for a random sample $X_1, ..., X_n \sim \text{Bernoulli}(p)$, the sufficient statistic is $\sum_{i=1}^{n}$. Why? Because taking the joint distribution, the sum of the variable values tells us everything we need to know for the likelihood function of p.

Because R would contain all the relevant information about S without changing anything about the parameter, we have $X \perp\!\!\!\perp S|R$.

If we have such a sufficient statistic R, then

$$\sum_r \Pr(Y \mid X=x, R=r)\Pr(R=r)$$
$$= \sum_{r,s} \Pr(Y, S=s \mid X=x, R=r)\Pr(R=r) \quad \text{(representing the marginal as a sum over the joint)}$$
$$= \sum_{r,s} \Pr(Y \mid X=x, R=r, S=s)\Pr(S=s \mid X=x, R=r)\Pr(R=r) \quad (f(X,Y)=f(X|Y)f(Y))$$
$$= \sum_{r,s} \Pr(Y \mid X=x, S=s)\Pr(S=s \mid X=x, R=r)\Pr(R=r) \quad (Y \perp\!\!\!\perp R|S)$$
$$= \sum_{r,s} \Pr(Y \mid X=x, S=s)\Pr(S=s \mid R=r)\Pr(R=r) \quad (X \perp\!\!\!\perp S|R)$$
$$= \sum_{s} \Pr(Y \mid X=x, S=s)\sum_r \Pr(S=s, R=r) \quad (f(X,Y)=f(X|Y)f(Y))$$
$$= \sum_{s} \Pr(Y \mid X=x, S=s)\Pr(S=s) \quad \text{(representing the sum over the joint as the marginal)}$$
$$= \Pr(Y \mid do(X=x))$$

But what is a relevant application of this result? Enter, the *propensity score*. A result presented here is that the propensity score defined as $f(S) = \Pr(X = 1|S = s)$ is a one-dimensional sufficient statistic that works in the event X is binary.

Knowing that $\Pr(R = r)$ is exclusively determined by the event $S = s$,

$$\sum_s \Pr(S = s)(E[Y \mid X = 1, S = s] - E[Y \mid X = 0, S = s])$$
$$= \sum_r \Pr(R = r)(E[Y \mid X = 1, R = r] - E[Y \mid X = 0, R = r])$$

Our ATE calculation has shifted from a problem where we regress $Y \sim X, S$ to one where we need to solve $X \sim S$ and use that result to get the much easier $Y \sim X, R$.

5 Identifying Causal Structure

The previous sections have discussed what we can do if we have access to the DAG in question. However, this may not always be the case. Another part of causal inference is actually identifying the causal graph itself.

We outline 2 starting strategies for identifying causal structure and leave the advanced reader to explore the discovery algorithms.

5.1 Prior Information

Prior information is relatively straightforward when describing graph structure. If you know from previous information that an edge in a causal graph exists, you should use it. While this is not statistical, it is important to mention because prior information can impose a set of constraints on the graph which must be abided by in further analysis.

5.2 Guessing and Testing

5.2.1 Testing DAGs

A presented DAG has 2 genres of claims in it. The first is direct causaion and the second is d-separation. We can test both of these claims.

If a DAG has an arrow such that $X \to Y$, if we create an experiment where we alter X alone and Y is unaltered, then the edge is false and should be removed.

Similarly, if we have a node (or set of nodes) S s.t. $X \perp\!\!\!\perp Y|S$, if we observe X, Y, and S and alter X for every value of S and see a relationship, then we know the d-separation claims is false and $X \not\perp\!\!\!\perp Y|S$.

5.2.2 Testing Conditional Independence

- If variables are discrete, one can just use a χ^2 test on a contingency table.
- If everything is linear multivariate Gaussian, then zero correlation implies independence.
- If the model is nonlinear, then we can use the definition of independence to see that $X \perp\!\!\!\perp Y|Z \rightarrow E(Y|Z) = E(Y|X,Z)$. So, if smoothing Y on Z leads to different predictions that smoothing Y on X and Z, then we do not have independence.

Note that faithfulness means that any conditional independence relationships we have MUST be expressed in the graph, but an issue can arise where different graphs can express the same conditional independence relationships (see pg. 470 for an example).

5.3 Discovery Algorithms

While I did not have the time to read throught this section in detail, there do exist algorithms that check the validity of the causal graphs.

- PC Algorithm
- SGS Algorithm