MS&E 226: Fundamentals of Data Science

Lecture 14: Introduction to causal inference

Ramesh Johari

Causation vs. association

Two examples

Suppose you are considering whether a new diet is linked to lower risk of inflammatory arthritis.

You observe that in a given sample:

- A small fraction of individuals on the diet have inflammatory arthritis.
- A large fraction of individuals not on the diet have inflammatory arthritis.

You recommend that everyone pursue this new diet, but rates of inflammatory arthritis are unaffected.

What happened?

Two examples

Suppose you are considering whether a new e-mail promotion you just ran is useful to your business.

You see that those who received the e-mail promotion did not convert at substantially higher rates than those who did not receive the e-mail.

So you give up...and later, another product manager runs an experiment with a similar idea, and conclusively demonstrates the promotion raises conversion rates.

What happened?

Association vs. causation

In each case, you were unable to see *what would have happened* to each individual if the alternative action had been applied.

- In the arthritis example, suppose only individuals predisposed to being healthy do the diet in the first place. Then you cannot see either what happens to an unhealthy person who *does* the diet, or a healthy person who *does not* do the diet.
- In the e-mail example, suppose only individuals who are unlikely to convert received your e-mail. Then you cannot see either what happens to an individual who is likely to convert who *receives* the promotion, or an individual who is not likely to convert who *does not receive* the promotion.

The lack of this information is what prevents inference about causation from association.

The “potential outcomes” model

Counterfactuals and potential outcomes

In our examples, the unseen information about each individual is the *counterfactual*.

Without reasoning about the counterfactual, we can't draw causal inferences—or worse, we draw the wrong causal inferences!

The *potential outcomes* model is a way to formally think about counterfactuals and causal inference.

Potential outcomes

Suppose there are two possible *actions* that can be applied to an individual:

- 1 ("treatment")
- 0 ("control")

(What are these in our examples?)

Potential outcomes

Suppose there are two possible *actions* that can be applied to an individual:

- 1 ("treatment")
- 0 ("control")

(What are these in our examples?)
For each individual in the population, there are *two* associated *potential outcomes*:

- $Y(1)$: outcome if treatment applied
- $Y(0)$: outcome if control applied

Causal effects

The *causal effect* of the action for an individual is the *difference* between the outcome if they are assigned treatment or control:

$$\text{causal effect} = Y(1) - Y(0).$$

The *fundamental problem of causal inference* is this:

> *In any example, for each individual, we only get to observe* one *of the two potential outcomes!*

In other words, this approach treats causal inference as a problem of *missing data*.

Assignment

The *assignment mechanism* is what decides which outcome we get to observe. We let $W = 1$ (resp., 0) if an individual is assigned to treatment (resp., control).

Assignment

The *assignment mechanism* is what decides which outcome we get to observe. We let $W = 1$ (resp., 0) if an individual is assigned to treatment (resp., control).

- In the arthritis example, individuals self-assigned.

Assignment

The *assignment mechanism* is what decides which outcome we get to observe. We let $W = 1$ (resp., 0) if an individual is assigned to treatment (resp., control).

- In the arthritis example, individuals self-assigned.
- In the e-mail example, we assigned them, but there was a bias in our assignment.

Assignment

The *assignment mechanism* is what decides which outcome we get to observe. We let $W = 1$ (resp., 0) if an individual is assigned to treatment (resp., control).

- In the arthritis example, individuals self-assigned.
- In the e-mail example, we assigned them, but there was a bias in our assignment.
- *Randomized* assignment chooses assignment to treatment or control at random.

Example 1: Potential outcomes

Here is a table depicting an extreme version of the arthritis example in the potential outcomes framework.

- $W = 1$ means the diet was followed
- $Y = 1$ or 0 based on whether arthritis was observed
- The *starred* entries are what we observe

Individual	W_i	$Y_i(0)$	$Y_i(1)$	Causal effect
1	1	0	0 $(*)$	0
2	1	0	0 $(*)$	0
3	1	0	0 $(*)$	0
4	1	0	0 $(*)$	0
5	0	1 $(*)$	1	0
6	0	1 $(*)$	1	0
7	0	1 $(*)$	1	0
8	0	1 $(*)$	1	0

Example 2: Potential outcomes

The same table can also be viewed as an extreme version of the e-mail example in the potential outcomes framework.

- $W = 1$ means the promotion was received
- $Y = 0$ means the individual converted; $Y = 1$ means the individual did not convert.
- The *starred* entries are what we observe

In each case the *association* is measured by examining the average difference of *observed* outcomes, which is 1. But the causal effects are all zero.

Mistakenly inferring causation

Suppose, e.g., in the arthritis data that you mistakenly infer causation, and encourage people to diet; half the non-dieters take up your suggestion.

Suppose you collect the same data again after this intervention:

Individual	W_i	$Y_i(0)$	$Y_i(1)$	Causal effect
1	1	0	0 $(*)$	0
2	1	0	0 $(*)$	0
3	1	0	0 $(*)$	0
4	1	0	0 $(*)$	0
5	1	1	1 $(*)$	0
6	1	1	1 $(*)$	0
7	0	1 $(*)$	1	0
8	0	1 $(*)$	1	0

Mistakenly inferring causation

Suppose, e.g., in the arthritis data that you mistakenly infer causation, and encourage people to diet; half the non-dieters take up your suggestion.

Suppose you collect the same data again after this intervention:

Individual	W_i	$Y_i(0)$	$Y_i(1)$	Causal effect
1	1	0	0 $(*)$	0
2	1	0	0 $(*)$	0
3	1	0	0 $(*)$	0
4	1	0	0 $(*)$	0
5	1	1	1 $(*)$	0
6	1	1	1 $(*)$	0
7	0	1 $(*)$	1	0
8	0	1 $(*)$	1	0

Now the average outcome among the non-dieters is still 1, while the average outcome among the dieters rises to 0.33: *conflating association and causation would suggest the intervention actually made things worse!*

Estimation of causal effects

“Solving” the fundamental problem

We can’t observe both potential outcomes for each individual.

So we have to get around it in some way. Some examples:

- Observe the same individual at different points in time
- Observe two individuals who are nearly identical to each other, and give one treatment and the other control

Both are obviously of limited applicability. What else could we do?

The average treatment effect

One possibility is to estimate the *average treatment effect* (ATE) in the population:

$$\mathsf{ATE} = \mathbb{E}[Y(1)] - \mathbb{E}[Y(0)].$$

In doing so we lose individual information, but now we have a reasonable chance of getting an estimate of both terms in the expectation.

Estimating the ATE

Let's start with the obvious approach to estimating the ATE:

- Suppose n_1 individuals receive the treatment, and n_0 individuals receive control.
- Compute:

$$\widehat{\mathsf{ATE}} = \frac{1}{n_1} \sum_{i:W_i=1} Y_i(1) - \frac{1}{n_0} \sum_{i:W_i=0} Y_i(0).$$

  Note that everything in this expression is observed.
- If both n_1 and n_0 are large, then (by LLN):

$$\widehat{\mathsf{ATE}} \approx \mathbb{E}[Y(1)|W=1] - \mathbb{E}[Y(0)|W=0].$$

The question is: when is this a good estimate of the ATE?

Selection bias

We have the following result.

Theorem

$\widehat{\text{ATE}}$ *is consistent as an estimate of the* ATE *if there is no* selection bias*:*

$$\mathbb{E}[Y(1)|W=1]=\mathbb{E}[Y(1)|W=0]; \;\; \mathbb{E}[Y(0)|W=1]=\mathbb{E}[Y(0)|W=0].$$

Selection bias

We have the following result.

Theorem
$\widehat{\text{ATE}}$ *is consistent as an estimate of the* ATE *if there is no* selection bias*:*

$$\mathbb{E}[Y(1)|W=1] = \mathbb{E}[Y(1)|W=0]; \;\; \mathbb{E}[Y(0)|W=1] = \mathbb{E}[Y(0)|W=0].$$

- In words: assignment to treatment should be uncorrelated with the outcome.
- This requirement is automatically satisfied if W is assigned randomly, since then W and the outcomes are *independent*. This is the case in a randomized experiment.
- It is *not* satisfied in the two examples we discussed.

Selection bias: Proof

Note that:

$$\begin{aligned}\mathbb{E}[Y(1)] &= \mathbb{E}[Y(1)|W=1]P(W=1)\\ &\quad + \mathbb{E}[Y(1)|W=0]P(W=0);\\ \mathbb{E}[Y(1)|W=1] &= \mathbb{E}[Y(1)|W=1]P(W=1)\\ &\quad + \mathbb{E}[Y(1)|W=1]P(W=0).\end{aligned}$$

Now subtract:

$$\begin{aligned}&\mathbb{E}[Y(1)] - \mathbb{E}[Y(1)|W=1] =\\ &\qquad \big(\mathbb{E}[Y(1)|W=0] - \mathbb{E}[Y(1)|W=1]\big)P(W=0).\end{aligned}$$

This is zero if the condition in the theorem is satisfied.

The same analysis can be carried out to show $\mathbb{E}[Y(0)] - \mathbb{E}[Y(0)|W=0] = 0$ if the condition in the theorem holds.

Putting the two terms together, the theorem follows.

The implication

Selection bias is rampant in conflating association and causation.

Remember to think carefully about selection bias in any causal claims that you read!

This is the reason why randomized experiments are the “gold standard” of causal inference: they remove any possible selection bias.

Randomized experiments

Randomization

In what we study now, we will focus on causal inference when the data is generated by a *randomized experiment*.[1]

In a randomized experiment, the assignment mechanism is random, and in particular independent of the potential outcomes.

How do we analyze the data from such an experiment?

[1] Other names: randomized controlled trial; A/B test

The estimator

Let's go back to $\widehat{\mathsf{ATE}}$:

$$\widehat{\mathsf{ATE}} = \frac{1}{n_1} \sum_{i:W_i=1} Y_i(1) - \frac{1}{n_0} \sum_{i:W_i=0} Y_i(0).$$

What is the variance of the sampling distribution of this estimator for a randomized experiment?

The estimator

Let's go back to $\widehat{\mathsf{ATE}}$:

$$\widehat{\mathsf{ATE}} = \frac{1}{n_1} \sum_{i:W_i=1} Y_i(1) - \frac{1}{n_0} \sum_{i:W_i=0} Y_i(0).$$

What is the variance of the sampling distribution of this estimator for a randomized experiment?

- For those i with $W_i = 1$, $Y_i(1)$ is an i.i.d. sample from the population marginal distribution of $Y(1)$.
  Suppose this has variance σ_1^2, which we estimate with the sample variance $\hat{\sigma}_1^2$ among the treatment group.

The estimator

Let's go back to $\widehat{\mathsf{ATE}}$:

$$\widehat{\mathsf{ATE}} = \frac{1}{n_1} \sum_{i:W_i=1} Y_i(1) - \frac{1}{n_0} \sum_{i:W_i=0} Y_i(0).$$

What is the variance of the sampling distribution of this estimator for a randomized experiment?

- For those i with $W_i = 1$, $Y_i(1)$ is an i.i.d. sample from the population marginal distribution of $Y(1)$.
  Suppose this has variance σ_1^2, which we estimate with the sample variance $\hat{\sigma}_1^2$ among the treatment group.
- For those i with $W_i = 0$, $Y_i(0)$ is an i.i.d. sample from the population marginal distribution of $Y(0)$.
  Suppose this has variance σ_0^2, which we estimate with the sample variance $\hat{\sigma}_0^2$ among the control group.

The estimator

Let's go back to $\widehat{\mathsf{ATE}}$:

$$\widehat{\mathsf{ATE}} = \frac{1}{n_1} \sum_{i:W_i=1} Y_i(1) - \frac{1}{n_0} \sum_{i:W_i=0} Y_i(0).$$

What is the variance of the sampling distribution of this estimator for a randomized experiment?

- For those i with $W_i = 1$, $Y_i(1)$ is an i.i.d. sample from the population marginal distribution of $Y(1)$.
  Suppose this has variance σ_1^2, which we estimate with the sample variance $\hat{\sigma}_1^2$ among the treatment group.
- For those i with $W_i = 0$, $Y_i(0)$ is an i.i.d. sample from the population marginal distribution of $Y(0)$.
  Suppose this has variance σ_0^2, which we estimate with the sample variance $\hat{\sigma}_0^2$ among the control group.
- So now we can estimate the variance of the sampling distribution of $\widehat{\mathsf{ATE}}$ as:

$$\widehat{\mathsf{SE}}^2 = \frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_0^2}{n_2}.$$

Asymptotic normality

For large n_1, n_0, the central limit theorem tells us that the sampling distribution fo $\widehat{\text{ATE}}$ is approximately normal:

- with mean ATE (because it is consistent when the experiment is randomized)
- with standard error $\widehat{\text{SE}}$ from the previous slide.

We can use these facts to analyze the experiment using the tools we've developed.

CIs, hypothesis testing, p-values

Using asymptotic normality, we can:

- Build a 95% confidence interval for ATE, as:

$$[\widehat{\mathsf{ATE}} - 1.96\widehat{\mathsf{SE}},\ \widehat{\mathsf{ATE}} + 1.96\widehat{\mathsf{SE}}].$$

- Test the null hypothesis that $\mathsf{ATE} = 0$, by checking if zero is in the confidence interval or not (this is the Wald test).
- Compute a p-value for the resulting test, as the probability of observing an estimate as extreme as $\widehat{\mathsf{ATE}}$ if the null hypothesis were true.

An alternative: Regression analysis

Another approach to analyzing an experiment is to use linear regression.

In particular, suppose we use OLS to fit the following model:

$$Y_i \approx \hat{\beta}_0 + \hat{\beta}_1 W_i.$$

In a randomized experiment, $W_i = 0$ or $W_i = 1$.

Therefore:

- $\hat{\beta}_0$ is the average outcome in the control group.
- $\hat{\beta}_0 + \hat{\beta}_1$ is the average outcome in the treatment group.
- So $\hat{\beta}_1 = \widehat{\text{ATE}}$!

We will have more to say about this approach next lecture.

An example in R

I constructed an "experiment" where $n_1 = n_0 = 100$, and:

$$Y_i = 10 + 0.5 \times W_i + \varepsilon_i,$$

where $\varepsilon_i \sim \mathcal{N}(0, 1)$. (Question: what is the true ATE?)

```
lm(formula = Y ~ 1 + W, data = df)
            coef.est coef.se
(Intercept) 9.9647   0.0953
W1          0.4213   0.1348
---
n = 200, k = 2
residual sd = 0.9532, R-Squared = 0.05
```

The estimated standard error on $\hat{\beta}_1 = \widehat{\text{ATE}}$ is the same as the estimated standard error we computed earlier.