MS&E 226: Fundamentals of Data Science

Lecture 13: 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).

- In the arthritis example, individuals self-assigned.
- In the e-mail example, we assigned them, but not completely at random.
- *Randomized* assignment chooses assignment to treatment or control at random.

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 not completely at random.
- *Randomized* assignment chooses assignment to treatment or control at random.

Note that we typically write Y for the *observed* outcome, i.e., $Y = Y(W)$.

(This is confusing notation because $Y(1), Y(0)$ are *potential* outcomes, and Y is *also* used for the *observed* outcome...but this is standard notation in causal inference.)

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 $(*)$	?
2	1	?	0 $(*)$	?
3	1	?	0 $(*)$	?
4	1	?	0 $(*)$	?
5	0	1 $(*)$	?	?
6	0	1 $(*)$	?	?
7	0	1 $(*)$	?	?
8	0	1 $(*)$	?	?

Causal inference as a missing data problem

The previous table makes it clear that the *missing data* prevents our ability to determine causal effects.

It is important that we wrote the table as on the previous slide! Usually, when we collect the data, it will just have one column for the observed outcome, instead of potential outcomes:

Individual	W_i	Y
1	1	0
2	1	0
3	1	0
...		

In this form, it masks the fact that causal inference depends on data that is in fact missing.

Example 1: Potential outcomes

Here is the table with the *missing data* filled in – we can't observe this!

- $W = 1$ 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	$Y_i(0)$	$Y_i(1)$	CE
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:

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

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:

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

  Note that everything in this expression is observed.

But is this an *unbiased* estimate of the ATE?

When does our estimator work?

If both n_1 and n_0 are large, then (by the law of large numbers):

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

But we want to estimate:

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

So our estimator works when:

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

When is this true?

No selection bias

We say there is *no selection bias* when:

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

No selection bias

We say there is *no selection bias* when:

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

In words:

- The average potential outcome *under treatment* is the same, regardless of whether the individual was assigned to treatment or control.
- The average potential outcome *under control* is the same, regardless of whether the individual was assigned to treatment or control.

No selection bias

We say there is *no selection bias* when:

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

In words:

- The average potential outcome *under treatment* is the same, regardless of whether the individual was assigned to treatment or control.
- The average potential outcome *under control* is the same, regardless of whether the individual was assigned to treatment or control.

When this holds, both treatment and control groups are *representative* of the overall population, and our estimator $\widehat{\text{ATE}}$ is unbiased for ATE. (See appendix for proof.)

Selection bias: Example

Arthritis example: Assignment to the diet was correlated with absence of arthritis:

- $\mathbb{E}[Y(0)|W = 1] = 0$ (those who chose the diet would not have had arthritis, even without the diet)
- $\mathbb{E}[Y(0)|W = 0] = 1$ (those who did not choose the diet had arthritis)

So the potential outcomes under control are *different* for those who chose the diet vs. those who did not $\implies$ *selection bias*.

(The same type of reasoning applies to the e-mail example: there the choice of who received the promotion led to selection bias.)

Confounding

When selection bias is present, we are comparing *different types of people*.

- Average outcome for treatment group under treatment $\approx \mathbb{E}[Y(1)|W = 1]$;
- Average outcome for control group under control $\approx \mathbb{E}[Y(0)|W = 0]$
- But if the groups are different, this difference reflects both:
  1. The causal effect of treatment, *and*
  2. The inherent differences between the two groups

When selection bias is present, we say that there is *confounding*: inherent differences between the two groups prevent us from precisely identifying the causal effect of treatment.

Confounding: A real-world example

In August 2021, as the Delta variant of COVID-19 was spreading, data from Israel was released:

- 18.2% of the population (1.3M) was not vaccinated; 78.7% of the population was vaccinated.
- 16.4 severe cases were observed per 100K *unvaccinated* individuals.
- 5.3 severe cases were observed per 100K *vaccinated* individuals.

This suggested a vaccine effectiveness of $1 - 5.3/16.4 \approx 67.5\%$, which was much lower than the reported 80-90% effectiveness of vaccines in trials...had the vaccine lost effectiveness in Israel?

Confounding by age

Two key facts about the Israeli population:

- *Age disparity in vaccination*:
  - 90.4% of residents > 50 years old were vaccinated (2.1M people).
  - Only 73% of residents < 50 years old were vaccinated (3.5M people).
- *Age disparity in severe disease risk*: Older people are at much higher risk of severe disease from COVID-19 than younger people.

In other words, through age, vaccination status (W) was *correlated* with severe disease risk given vaccination status ($Y(0), Y(1)$), creating a *selection bias*.

Confounding by age

When we stratify the data by age (and normalize per 100,000 people), we see a very different picture:

	Severe cases per 100k		Vaccine
Age group	Not Vax	Fully Vax	Effectiveness
< 50 years	3.9	0.3	91.8%
> 50 years	91.9	13.6	85.2%
Overall	16.4	5.3	67.5%

Within each age group, vaccines were in fact highly effective (85–92%).

In other words, the “perceived” drop in effectiveness could be almost entirely explained by confounding.

(Example adapted from Jeffrey Morris, *Covid Data Science* blog, August 17, 2021.)

Simpson's paradox

This form of selection bias is called *Simpson's paradox*:

When a confounding variable is present, the association observed in the overall population can be substantially different (or even *reversed*) compared to the true causal effects when we stratify by the confounding variable.

The key issue here is that we are trying to use *observational* data to study causal effects, i.e., data where we don't control assignment.

Confounding due to selection bias is a key challenge in causal inference from observational data; we will discuss this later in the course.

Another example: Berkeley admissions

Berkeley was sued for gender bias in admissions to graduate school based on 1973 statistics: 44% of men were admitted, while only 35% of women were admitted.

But based on individual departments’ admissions statistics, there did not appear to be statistically significant gender-based discrimination (in fact if anything, some departments tended to *favor* women).

The evidence in the case revealed an unusual example of Simpson’s paradox: women were systematically applying to majors that were much more competitive, creating a selection bias.

Randomized experiments

Randomization eliminates selection bias

If assignment W is *randomized* independently across units, then W is independent of the potential outcomes $Y(0)$ and $Y(1)$.

This means:

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

There is no selection bias, and our estimator $\widehat{\text{ATE}}$ is *unbiased*.

This is why randomized experiments are the "gold standard" for causal inference.

The completely randomized design

We now 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.

Specifically, we study a *completely randomized design* (CRD): we randomly assign units to treatment and control, but constrain the total number of treatment units to be fixed at n_1, and the total number of control units to be fixed at n_0.

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 $\widehat{\mathrm{ATE}}$:

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

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

The estimator

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

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

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 $Y(1)$, with variance σ_1^2 (estimated by sample variance $\hat{\sigma}_1^2$).
- For those i with $W_i = 0$, $Y_i(0)$ is an i.i.d. sample from the population marginal distribution of $Y(0)$, with variance σ_0^2 (estimated by sample variance $\hat{\sigma}_0^2$).
- The variance of the sampling distribution of $\widehat{\mathrm{ATE}}$ is estimated as:

$$\widehat{\mathrm{SE}}^2 = \frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_0^2}{n_0}.$$

Asymptotic normality

For large n_1, n_0, the central limit theorem tells us that the sampling distribution of $\widehat{\text{ATE}}$ 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.

Confidence intervals, hypothesis testing, p-values

Using asymptotic normality, we can:

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

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

- Test the null hypothesis that ATE $= 0$, by checking if zero is in the confidence interval or not.
- Compute a p-value for the resulting test, as the probability of observing a t statistic as extreme as $|\widehat{\text{ATE}}/\widehat{\text{SE}}|$ if the null hypothesis were true.

SUTVA and interference [∗]

SUTVA and interference [$*$]

Interference [$*$]

Implicitly throughout our discussion of causal inference, we have assumed there is no *interference* between treatment and control:

Whether or not individual i receives treatment or control has *no impact* on the causal effect of treatment on another individual j.

When might this fail?

Interference [$*$]

Suppose Airbnb decides to A/B test a new feature that dramatically simplifies the booking process for a guest.

In the test, guests are randomized at when they start the booking process; control is the old experience, treatment is the new experience.

It is found that customers with the new experience book much more frequently than customers with the old experience, but the estimated $\widehat{\mathrm{ATE}}$ is an *overestimate*. Why?

Interference [$*$]

Both treatment and control see the *same* inventory of host listings!

So if treatment individuals book more often, that *reduces* the inventory available to control individuals, and implies their booking rates will be lower.

SUTVA [$*$]

If interference is present, the “potential outcomes” for an individual are much more complicated: they depend on not just the treatment a single individual received, but also on the treatment *other* individuals received.

With n individuals, this is 2^n potential outcomes for each individual!

The assumption that there is no interference between treatment and control is part of the *stable unit treatment value assumpton* (SUTVA) in econometrics and causal inference.

(The other part of SUTVA is that there is only one form of treatment or control: e.g., if treatment is “taking a drug”, there should be no variation in the treatment group as to *how much* of the drug is taken.)

Appendix [*]

Formal theorem on selection bias [$*$]

Theorem

Assume that observations are independently drawn from the population. Then if there is no selection bias, $\widehat{\text{ATE}}$ *is unbiased as an estimate of the* ATE*:*

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

In other words: the estimator $\widehat{\text{ATE}}$ is unbiased when assignment to treatment is uncorrelated with the potential outcomes.

Proof of the theorem [$*$]

Let $\mathbf{W} = (W_1, \dots, W_n)$ denote the assignments of the n observations.

Let $n_1 = \sum_{i=1}^{n} W_i$ and $n_0 = n - n_1$ be the number of units assigned to treatment and control, respectively.

With no selection bias, for each i:

$$\mathbb{E}[Y_i(1)|\mathbf{W}] = \mathbb{E}[Y_i(1)|W_i] = \mathbb{E}[Y_i(1)]; \quad \mathbb{E}[Y_i(0)|\mathbf{W}] = \mathbb{E}[Y_i(0)|W_i] = \mathbb{E}[Y_i(0)] \text{ for all } i.$$

Therefore, conditional on the assignments $\mathbf{W}$:

$$\begin{aligned}\mathbb{E}[\widehat{\text{ATE}}|\mathbf{W}] &= \frac{1}{n_1} \sum_{i:W_i=1} \mathbb{E}[Y_i(1)|\mathbf{W}] - \frac{1}{n_0} \sum_{i:W_i=0} \mathbb{E}[Y_i(0)|\mathbf{W}] \\ &= \frac{1}{n_1} \sum_{i:W_i=1} \mathbb{E}[Y_i(1)] - \frac{1}{n_0} \sum_{i:W_i=0} \mathbb{E}[Y_i(0)].\end{aligned}$$

Proof of the theorem (continued) [$*$]

Since units are drawn independently from the population, we have $\mathbb{E}[Y_i(1)] = \mathbb{E}[Y(1)]$ and $\mathbb{E}[Y_i(0)] = \mathbb{E}[Y(0)]$ for all i.

Therefore:

$$\begin{aligned}\mathbb{E}[\widehat{\text{ATE}}|\mathbf{W}] &= \frac{1}{n_1}\sum_{i:W_i=1}\mathbb{E}[Y(1)] - \frac{1}{n_0}\sum_{i:W_i=0}\mathbb{E}[Y(0)] \\ &= \mathbb{E}[Y(1)] - \mathbb{E}[Y(0)] = \text{ATE}.\end{aligned}$$

Since this holds for *any* assignment pattern $\mathbf{W}$:

$$\mathbb{E}[\widehat{\text{ATE}}] = \mathbb{E}_{\mathbf{W}}[\mathbb{E}[\widehat{\text{ATE}}|\mathbf{W}]] = \mathbb{E}_{\mathbf{W}}[\text{ATE}] = \text{ATE}.$$

Thus $\widehat{\text{ATE}}$ is unbiased when there is no selection bias.