MA50259: Statistical Design of Investigations

Dr. Sandipan Roy

Lecture 7: Observational studies and Causal Inference

Observational studies

- An observational study is an empirical investigation of the relation between exposures and outcomes (or treatment and effects) when randomised experimentation is unethical or infeasible
- A well designed observational study resembles, as closely as possible, a simple randomised experiment
- Unlike experiments, in some observational studies, the outcomes may exist as measurements prior to the design of the study; it is their examination and use, not their existence, that separates design from analysis

Causal inference

- Causal inference is the study of inferring the presence and magnitude of cause-effect relationships from data
- As sociologists, economists, epidemiologists etc., and indeed as human beings, it is something we know an awful lot about
- Suppose a study finds an association between government spending in technology and suicide by hanging
- On the back of this, the government cuts the budget on technology with a view to reducing suicide by hanging
- We would all agree that this is silly!
- This is because we understand the difference between association and causation

Simple example

- 12 individuals each suffer a headache
- Some take a pill; others don't
- One hour later, we ask each of the 12 whether or not his/her headache has disappeared

The observed data

Here are the data

	X pill taken?	Y headache dissapeared?
Jose	0	0
Pancho	1	0
Juan	1	1
Miguel	0	0
Maria	0	1
Rosa	1	0
Gloria	1	0
Manuel	0	0
Jorge	0	1
Guadalupe	0	0
Carmen	0	1
Leticia	1	1

The observed data

	X pill taken?	Y headache dissapeared?
Jose	0	0
Pancho	1	0
Juan	1	1
Miguel	0	0
Maria	0	1
Rosa	1	0
Gloria	1	0
Manuel	0	0
Jorge	0	1
Guadalupe	0	0
Carmen	0	1
Leticia	1	1

- Juan took the pill, and his headache disappeared
- Did the pill cause his headache to disappear?
- We do not know
- To answer this, we need to know what would have happened had he not taken the pill

Counterfactuals and potential outcomes

- X is the treatment: whether or not a pill was taken
- Y is the outcome: whether or not the headache disappeared
- Write Y^0 and Y^1 to represent the potential outcomes under both treatments.
- Y^0 is the outcome which would have been seen had the pill NOT been taken.
- Y^1 is the outcome which would have been seen had the pill been taken.
- One of these is observed: if $X = 0$, Y^0 is observed; if $X = 1$, Y^1 is observed.
- The other is counterfactual

The ideal data (1)

Suppose that we can observe the unobservable

	Y^0	Y^1
Jose	0	0
Pancho	1	0
Juan	0	1
Miguel	0	0
Maria	1	1
Rosa	0	0
Gloria	0	0
Manuel	0	0
Jorge	1	0
Guadalupe	0	0
Carmen	1	1
Leticia	0	1

- For Juan, the pill did have a causal effect
- He did take it, and his headache disappeared; but had he not taken it, his headache would not have disappeared
- Thus the pill had a causal effect on his headache
- what about Rosa?
- and Jorge?

The ideal data (2)

Suppose that we can observe the unobservable

	Y^0	Y^1	Causal effect?
Jose	0	0	No
Pancho	1	0	Yes,harmful
Juan	0	1	Yes,protective
Miguel	0	0	No
Maria	1	1	No
Rosa	0	0	No
Gloria	0	0	No
Manuel	0	0	No
Jorge	1	0	Yes,harmful
Guadalupe	0	0	No
Carmen	1	1	No
Leticia	0	1	Yes,protective

- An individual-level causal effect is defined for each subject and is given by $Y^1 - Y^0$
- These need not all be the same.

The fundamental problem of causal inference

Back to reality

	Y^0	Y^1	X	Y
Jose	0	?	0	0
Pancho	?	0	1	0
Juan	?	1	1	1
Miguel	0	?	0	0
Maria	1	?	0	1
Rosa	?	0	1	0
Gloria	?	0	1	0
Manuel	0	?	0	0
Jorge	1	?	0	1
Guadalupe	0	?	0	0
Carmen	1	?	0	1
Leticia	?	1	1	1

- In reality, we never observe both Y^0 and Y^1 on the same individual
- Sometimes called the fundamental problem of causal inference
- It is therefore over-ambitious to try to infer anything about individual-level causal effects.

Population-level causal effects (1)

- A less ambitious goal is to focus on the population-level or average causal effect:

$$E[Y^1] - E[Y^0]$$

or, since Y is binary,

$$P(Y^1 = 1) - P(Y^0 = 1)$$

- Lets return to the "ideal" data. . .

Population-level causal effects (2)

	Y^0	Y^1	Causal effect?
Jose	0	0	No
Pancho	1	0	Yes,harmful
Juan	0	1	Yes,protective
Miguel	0	0	No
Maria	1	1	No
Rosa	0	0	No
Gloria	0	0	No
Manuel	0	0	No
Jorge	1	0	Yes,harmful
Guadalupe	0	0	No
Carmen	1	1	No
Leticia	0	1	Yes,protective

$$P(Y^0 = 1) = \frac{4}{12}$$

$$P(Y^1 = 1) = \frac{4}{12}$$

$$P(Y^1 = 1) - P(Y^0 = 1) = 0$$

i.e. no causal effect at the population level

Population-level causal effects (3)

- In reality, we dont know Y^1 for every subject, so we cant simply estimate $P(Y^1 = 1)$ as the proportion of all subjects with $Y^1 = 1$.
- Likewise, we cant simply estimate $P(Y^0 = 1)$ as the proportion of all subjects with $Y^0 = 1$.
- Thus we cant easily estimate $P(Y^1 = 1) - P(Y^0 = 1)$
- Causal inference is all about choosing quantities from the observed data (i.e. involving X , Y and other observed variables) that represent reasonable substitutes for hypothetical quantities such as $P(Y^1 = 1) - P(Y^0 = 1)$, which involve unobservable counterfactuals

Possible Solutions

Substitutes: Cannot apply two treatments to the same unit at the same time but can

- Apply treatments at different times
- Subdivide the unit
- Use the pre-treatment condition as the control

All require strong assumptions!

When does association = causation? (1)

- What might be a good substitute for $P(Y^1 = 1)$?
- What about $P(Y = 1|X = 1)$?
- This is the proportion whose headache disappeared among those who actually took the pill.
- Is this the same as $P(Y^1 = 1)$?
- Only if those who took the pill are exchangeable with those who didnt. $(Y^0, Y^1) \perp X$ (Ignorability)
- This would be the case if the choice to take the pill was made at random
- This is why ideal randomised experiments are the gold standard for inferring causal effects

Randomization

- Cannot apply treatments to the same unit but can apply to similar units
- Randomization of the treatment is used to obtain the exchangeabilty
- Require the idea of using a sample to learn about a population. Inference is about the population not individuals
- Also requires assumptions and may not be practical

When does association = causation? (2)

	Y^0	Y^1	X	Y
Jose	0	?	0	0
Pancho	?	0	1	0
Juan	?	1	1	1
Miguel	0	?	0	0
Maria	1	?	0	1
Rosa	?	0	1	0
Gloria	?	0	1	0
Manuel	0	?	0	0
Jorge	1	?	0	1
Guadalupe	0	?	0	0
Carmen	1	?	0	1
Leticia	?	1	1	1

$$P(Y = 1|X = 1) = \frac{2}{5}$$

$$P(Y = 1|X = 0) = \frac{3}{7}$$

$$P(Y = 1|X = 1) -$$

$$P(Y = 1|X = 0) = -\frac{1}{35}$$

If we assumed that association = causation, we would conclude that the pill was, on average, slightly harmful

What's going on?

	Y^0	Y^1	X	Y
Jose	0	0	0	0
Pancho	1	0	1	0
Juan	0	1	1	1
Miguel	0	0	0	0
Maria	1	1	0	1
Rosa	0	0	1	0
Gloria	0	0	1	0
Manuel	0	0	0	0
Jorge	1	0	0	1
Guadalupe	0	0	0	0
Carmen	1	1	0	1
Leticia	0	1	1	1

- The subjects with the more severe headaches are more likely to take the pill
- So association $\neq$ causation.

Observational studies

- Randomization means that treatment groups are balanced
- Observational study means that treatments are observed not assigned
- Systematic differences may exist between the treatment groups
- Confounding variables may be associated with the response and the membership of the treatment groups
- Now more difficult (but not impossible) to estimate the causal effect

Taking severity into account

- Suppose we asked each of the 12 subjects at the beginning of the study: is your headache severe?.
- Then, we might propose that, after taking severity into account, the decision as to whether or not to take the pill was effectively taken at random
- Suppose Z denotes severity. Then, under this assumption, within strata of Z , the exposed and unexposed subjects are exchangeable
- This is called conditional exchangeability (given Z).
  $(Y^0, Y^1) \perp X|Z$ (ignorability)
- Under conditional exchangeability given Z , association $=$ causation within strata of Z .
- Lets return to the data and look for an association between X and Y within strata of Z

Stratifying on severity

	Y^0	Y^1	X	Y	Z
Jose	0	0	0	0	1
Pancho	1	0	1	0	0
Juan	0	1	1	1	0
Miguel	0	0	0	0	1
Maria	1	1	0	1	0
Rosa	0	0	1	0	1
Gloria	0	0	1	0	1
Manuel	0	0	0	0	0
Jorge	1	0	0	1	1
Guadalupe	0	0	0	0	0
Carmen	1	1	0	1	0
Leticia	0	1	1	1	1

In the stratum $Z = 0$

$$P(Y = 1|X = 1) = \frac{1}{2}$$

$$P(Y = 1|X = 0) = \frac{2}{4}$$

In the stratum $Z = 1$

$$P(Y = 1|X = 1) = \frac{1}{3}$$

$$P(Y = 1|X = 0) = \frac{1}{3}$$

i.e. within strata of Z we find no association between X and Y

Summary so far (1)

- We have looked at a simple, artificial example, and defined what we mean by a causal effect
- We have seen that, unless the exposed and unexposed groups are exchangeable, association is not causation
- In our simple example, there was no (average) causal effect of X on Y .
- And yet, X and Y were associated, because of Z .

Summary so far (2)

- When we stratified on Z , we found no association between X and Y .
- So association = causation within strata of Z .
- This is because exposed and unexposed subjects were conditionally exchangeable given Z
- More generally, when there is a causal effect of X on Y , but also a non-causal association via Z , the causal effect will be estimated with bias unless we stratify on Z .

Summary so far (3)

- Conditional exchangeability is the key criterion that allows us to make causal statements using observational data.
- We need to identify, if possible, a set of variables $Z_1, Z_2, \ldots$, such that conditional exchangeability holds given these.
- In real life, there may be many candidate Z -variables.
- These may be inter-related in a very complex way.
- Deciding whether or not the exposed and unexposed are conditionally exchangeable given $Z_1, Z_2, \ldots$ requires detailed background subject-matter knowledge.

Summary so far (4)

- Causal diagrams can help us to use this knowledge to determine whether or not conditional exchangeability holds
- we may find it thoroughly implausible that we have collected data on a sufficient set of confounders such that conditional exchangeability holds.
- In this situation, we can sometimes still make causal inferences, if we are prepared to make an alternative set of assumptions and this points towards an alternative set of statistical methods.

- Suppose we have (carefully) identified a sufficient set of variables to control for confounding. What next?
- If the number of confounders is small and categorical/binary, we could stratify on them. We would then calculate our effect of interest (e.g. an odds ratio) in each stratum and then combine these in the usual way (MantelHaenszel), or report them separately if there are effect modifiers of interest, etc.
- Or, if there are too many confounders and/or some are continuous, we could specify a suitable regression model (linear/logistic/Poisson/Cox. . .):

$$\mathit{logit}\, Pr(D = 1|E, \mathit{gender}, \mathit{age}) = \alpha + \beta E + \gamma\, \mathit{gender} + \delta\, \mathit{age}$$

- There are other options. Instead of modelling the outcome given the exposure and confounders, we can model the outcome given the exposure (only), AND model the exposure given the confounders (and then stratify, match or re-weight using predictions from this model). Such methods are called propensity score methods.
- Other option is matching (for example in case-control studies) More later!

Example

Question: Does eating dark green vegetables protect against stomach cancer?

- Fixed term cohort study: 752 cancers out of 10,000 female nurses (aged 30-70y)
- Exposure: usual intake at baseline
- Potential confounders: age, smoking, alcohol intake

Table: Odds ratio (OR) of dark green vegetable intake

Adjusted for:	OR	95%CI
Age	1.04	(0.97,1.11)
Age + alc	1.52	(1.40,1.66)
Age + smok	1.03	(0.96,1.11)
Age+alc+smok	1.39	(1.27,1.52)

Which OR to report (assuming no interactions)? Interpretations:

1. Among those of a particular age, smoking & alcohol status: the expected percentage increase in cancer odds were everybody to add 1 unit to their dark green vegetable intake? (association)
2. Under ignorability assumptions we can make this interpretation causal!