Introduction to Causal Inference

Ashley I. Naimi, PhD

Outline

Causal Inference

- Introduction
- The Logic of Causal Inference
- Complex Longitudinal Data
- Notation
- Estimand, Estimator, Estimate
- Identifiability
  a. Counterfactual Consistency
  b. No Interference
  c. Excheangability
  d. Correct Model Specification
  e. Positivity
- Final Points

Causal Inference

Introduction

"Causal inference" deals primarily with the formal mechanisms by which we can combine data, assumptions, and models to interpret a correlation (or association) as a causal relation.[1] The framework by which we define what we mean by "causal relation" or "causal effect" is the **potential outcomes framework**.

A central notion in the potential outcomes framework is the counterfactual. This notion stems from the intuitive and informal practice of interpreting cause-effect relations as **circumstances (e.g., health outcomes) that would have arisen had things (e.g., exposures) been different**.

While this intuition serves an important purpose, it is not sufficient for doing rigorous science. Suppose we ask: "what is the effect of smoking on CVD risk, irrespective of smoking's effect on body weight?" This question seems clear and intiutive. To answer this question, we would do a study in which we collect data, enter these into a computer, perform some calculations, and obtain a number (the "effect").

But there is a problem.[2] The calculations performed by the computer are **rigorously defined mathematical objects**. On the other hand, **english language sentences about cause effect relations are ambiguous**. For example, the "effect of smoking" can mean many different things:

- All people smoke any tobacco ever versus no people smoke tobacco ever.
- All people smoke 3 cigarettes per day versus all people smoke 2 cigarettes per day.
- All people who have smoked any tobacco in the last 15 years cease to smoke any tobacco whatsoever.

Similarly, "irrespective of" can mean a number of things:

- The effect of smoking on CVD risk that would be observed in a hypothetical world where smoking did not affect body mass?

[1] There are a number of excellent introductory books and articles on causal inference in the empirical sciences. Here are some excellent options: M. A. Hernán and Robins (Forthcoming), Pearl, Glymour, and Jewell (2016), Imbens and Rubin (2015)

[2] This problem was articulated by Robins (1987), and I am using the example from his paper.

- The effect of smoking on CVD risk if everyone were set to "normal" body mass?
- The effect of smoking on CVD risk if everyone were held at the body mass they had in the month prior to study entry?

But the numerical strings of data and the computer algorithms applied to these data are well defined mathematical objects, which do not admit such ambiguity. Depending on several choices, including the data, how variables are coded, and the modeling strategy, the computer is being told which question to answer. There is a lot of potential uncertainty in the space between the English language sentences we use to ask causal questions, and the computer algorithms we use to answer those questions. Causal inference is about clarifying this uncertainty.

The Logic of Causal Inference

Recently, some authors have raised concerns about the increasing popularity of causal inference methods (e.g., Vandenbroucke, Broadbent, and Pearce 2016,Krieger and Davey Smith (2016)). Many of the papers written on this topic share a common objection: inferring causality based on a set of methods that require heroic assumptions will ultimately constrain our ability to improve population health.

Some of these papers make some important points.[3] However, they are premised on a fundamental misunderstanding of the logic of causal inference. In particular, these papers presume that the "causal inference" approach to inferring causality proceeds as follows:

> If our assumptions hold, then we can interpret the estimated association causally.

Rather, loosely speaking, the general structure that "causal inference" gives is the following:

> For a given dataset and a given target parameter, then here are the assumptions we need to interpret the estimated association causally.

There are many issues (some subtle) that I'm ignoring here. However, in the remainder of this course, I will try to clarify precisely

[3] In an excellent response to these concerns, Greenland (2017) recently pointed out that there are in fact issues with the "causal inference" framework, notably it's name. He prefers "causal modeling."

what this general structure of causal inference is, and why the distinction between these two ways of framing causal inference is of critical importance.

Complex Longitudinal Data

This short course is about methods that can be applied to measured at a single time point, as well as complex longitudinal data. For clarity, let's define complex longitudinal data. We will be dealing with data from a cohort study, individuals sampled from a well-defined target population, and clear study start and stop times (i.e., closed cohort). Data from such a cohort are **longitudinal** when they are measured repeatedly over time.[4]

[4] Another such form is when data are measured repeatedly across space. We will not be dealing with these data here.

Different scenarios can lead to longitudinal data:

1. exposure and covariates do not vary over time, but the study outcome can occur more than once
2. exposure and covariates vary over time, but the study outcome can only occur once
3. exposure and covariates vary over time, and the study outcome can occur more than once

We will deal with data that from scenario 2 (however, it is not difficult to generalize the logic to scenario 3). Repeated exposure, covariate, and/or outcome measurement is what leads to "longitudinal" data. But why complex?

Repeated measurement over time opens up the possibility of complex causal relations between past and future covariates. Suppose we measure an expsoure twice over follow-up, a covariate once, and the outcome at the end of follow-up (Figure 1). If we can assume that past exposure/covariate values do not affect future exposure/covariate values (usually a very risky assumption), we might not consider these data "complex," becuase we can use many standard methods we already know to analyze these data.

On the other hand, if past exposure/covariates affect future exposure/covariates in such a way that prior exposures or covariates confound future exposures (Figure 2), more advanced analytic techniques are needed.

Figure 1: Longitudinal data that might not be considered 'complex' because there is no feedback between exposure and covariates.

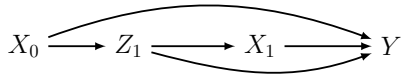

Figure 2: The simplest kind of complex longitudinal data. Note that the exposure at time zero affects the covariate at time 1 which affects the exposure at time 1. This feedback leads to confounding of the time 1 expsoure by a covariate that is affected by the prior exposure. Analysis of these data require more general methods to account for this complex form of confounding.

In this short course, we will learn how to use g methods to account for this type of complex time-varying confounding.

Notation

The building blocks for causal inference are **potential outcomes** (Rubin 2005). These are conceptually distinct from **observed outcomes**. Potential outcomes are functions of exposures. For a given exposure x, we will write the potential outcome as Y^x.[5] **This is interpreted as "the outcome (Y) that would be observed if X were set to some value x"**. For example, if X is binary [denoted $X \in (0,1)$], then Y^x is the outcome that would be observed if $X = 0$ or $X = 1$. If we wanted to be specific about the value of x, we could write $Y^{x=0}$ or $Y^{x=1}$ (or, more succinctly, Y^0 or Y^1).

[5] Alternate notation includes: Y_x, $Y(x)$, $Y \mid Set(X = x)$, and $Y|do(X = x)$.

Study Question 1: Suppose you collect data from a single person and find that they are exposed. Can you interpret their outcome to be the potential outcome that would have been observed had they been exposed? Why or why not?

When the exposoure and/or outcome are measured repeatedly over follow-up, notation must account for that. We thus use subscripts to denote when the variable was measured. For example, if the exposure is measured twice, we can denote the first measurement X_0 and the second X_1. Additionally, we use overbars to denote

the history of a variable over follow-up time. For example, $\overline{X}_1$ denotes the set $\{X_0, X_1\}$. More generally, for some arbitrary point over follow-up m, $\overline{X}_m$ denotes $\{X_0, X_1, X_2, \ldots X_m\}$. We can then define potential outcomes as a function of these exposure histories: For two exposure measurements, $\overline{X}_j = \{1,1\}$, $Y^{\overline{x}_j=\overline{1}}$ is the outcome that would be observed if X_0 were set to 1 and X_1 were set to 1.

Estimand, Estimator, Estimate

Causal inference starts with a clear idea of the effect of interest (the target causal parameter). To do this, it helps to distinguish between estimands, estimators, and estimates.

STUDY QUESTION 2A: You are familar with the well known odds ratio equation for a 2×2 table: (ab/cd). Is this an estimand, estimator, or estimate?

The **estimand** is the (mathematical) object we want to quantify. It is, for example, the causal risk difference, risk ratio, or odds ratio for our exposure and outcome of interest. In our smoking CVD example, we might be interested in:

$$E(Y^1 - Y^0), \quad \frac{E(Y^1)}{E(Y^0)}, \quad \frac{Odds(Y^1 = 1)}{Odds(Y^0 = 1)},$$

where $Odds(Y^x = 1) = E(Y^x)/[1 - E(Y^x)]$, and where $E(.)$ is the expectation operator taken with respect to the total population.[6] There are many others besides these.

[6] Throughout this course, if the outcome Y is binary, then $E(Y) \equiv P(Y = 1)$. Or, the expectation of Y is equivalent to the probability that $Y = 1$. For the more technially oriented,

$$E(Y) = \int yf(y)dy$$

where $f(y)$ is the probability density function of Y.

STUDY QUESTION 2B: List some estimators that can be used to quantify the odds ratio.

The estimand is the object we want to estimate. The **estimator** is an equation that allows us to use our data to quantify the estimand. Suppose, for example, we were explicitly intersted in quantifying

the causal risk difference for the relation between smoking and CVD risk. To do this, we have to start by quantifying the associational risk difference, but there are many ways to do this, including ordinary least squares, maximum likelihood, or the method of moments.

To be specific, let's simulate some hypothetical data on the relation between smoking and CVD. Let's look at ordinary least squares and maximum likelihood as estimators:

```
### CODE SET 1
# define the expit function
expit<-function(z){1/(1+exp(-(z)))}
set.seed(123)
n<-1e6
confounder<-rbinom(n,1,.5)
smoking<-rbinom(n,1,expit(-2+log(2)*confounder))
CVD<-rbinom(n,1,.1+.05*smoking+.05*confounder)

round(mean(confounder),3)

## [1] 0.499

round(mean(smoking),3)

## [1] 0.166

round(mean(CVD),3)

## [1] 0.133

#OLS
round(coef(lm(CVD~smoking+confounder)),4)

## (Intercept)     smoking  confounder
##      0.1000      0.0485      0.0501

#ML1
round(coef(glm(CVD~smoking+confounder,family=poisson("identity"))),4)

## (Intercept)     smoking  confounder
##      0.0999      0.0487      0.0502
```

```
#ML2
round(coef(glm(CVD~smoking+confounder,family=binomial("identity"))),4)

## (Intercept)     smoking  confounder
##      0.1000      0.0487      0.0501

### END CODE SET 1
```

In our simple setting with 1 million observations, ordinary least squares and maximum likelihood yielded the same associational risk difference (as expected) even though they are different **estimators**. Finally, the values obtained from each regression approach are our **estimates**.

Identifiability

In our simulation example, we estimated the associational risk difference using three different estimators. Estimating associations is all we can do with empirical data. But we want to use the associational risk difference to quantify the causal risk difference. We can only do so if the causal risk difference is **identified**. *A parameter (e.g., causal risk difference) is identified if we can write it as a function of the observed data.*

The causal risk difference is defined as a contrast of potential outcomes. Referring back to our simulated example, we want to estimate the causal risk difference which is an example of an average treatment effect:

$$E(Y^1 - Y^0),$$

where Y^1, Y^0 are the potential CVD outcomes that would be observed if smoking were set to 1 and 0, respectively. On the other hand, the associational risk difference is defined as a contrast of observed outcomes:

$$E(Y \mid X = 1) - E(Y \mid X = 0),$$

where each term in this equation is interpreted as the risk of CVD **among those who had** $X = x$. The causal risk difference is identified if the following equation holds:[7]

$$E(Y^x) = E(Y \mid X = x)$$

[7] Throughout this course, we will assume that the target parameter of interest is a causal contrast of potential outcomes. Sometimes, the target parameter of interest is an associational contrast, and the assumptions needed are less demanding. See, e.g., Naimi et al. (2016).

which says that the risk of CVD that would be observed if everyone were set to $X = x$ is equal to the risk of CVD that we observe among those with $X = x$. In this equation, the right hand side equation is written entirely in terms of observed data ($Y = 1$). The left hand side is a function of unobserved potential outcomes ($Y^x = 1$). This equivalence will only hold if we can make some assumptions.

The first is **counterfactual consistency**, which states that the potential outcome that would be observed if we set the exposure to the observed value is the observed outcome (Miguel A. Hernán 2005,Hernan and Taubman (2008),Miguel A Hernán and VanderWeele (2011),VanderWeele and Hernán (2013)).[8] Formally, counterfactually consistency states that:

$$\text{if } X = x \text{ then } Y^x = Y$$

The status of this assumption remains unaffected by the choice of analytic method (e.g., standard regression versus g methods). Rather, this assumption's validity depends on the nature of the exposure assignment mechanism.

We must also assume **no interference**, which states that the potential outcome for any given individual does not depend on the exposure status of another individual (M. G. Hudgens and Halloran 2008,Naimi and Kaufman (2015)). If this assumption were not true, we would have to write the potential outcomes as a function of the expsoure status of multiple individuals. For example, for two different people indexed by i and j, we might write: $Y_i^{x_i,x_j}$.[9] Notation and methods that account for interference can be somewhat complex (Tchetgen Tchetgen and VanderWeele 2012,M. E. Halloran and Hudgens (2016)), and we will not consider the impact of interference here.

Together, counterfactual consistency and no interference allow us to make some progress in writing the potential risk $E(Y^x)$ as a function of the observed risk $E(Y \mid X = x)$. Specifically, by counterfactual consistency and no interference, we can do the following:

$$E(Y \mid X = x) = E(Y^x \mid X = x)$$

[8] While somewhat convoluted, this assumption is about legitimizing the connection between our observational study, and future interventions in actual populations. In our observational study, we **see** people with with a certain value of the exposure. In a future intervention, we **set** people to a certain value of the exposure.

[9] Together, counterfactual consistency and no interference make up the stable-unit treatment value assumption (SUTVA), first articulated by D. B. Rubin (1980).

A third assumption is **exchangeability**, which implies that the potential outcomes under a specific exposure (Y^x) are independent of the observed exposures X (Greenland and Robins 1986,Greenland, Robins, and Pearl (1999),Greenland and Robins (2009)). If this holds, then we have:

$$E(Y^x \mid X = x) = E(Y^x)$$

If there is any confounding, selection, or information bias, the potential outcome will be associated with the observed exposure, and we cannot remove $X = x$ from the conditioning statement.[10] What this means is that the exposure is predictive of prognosis, independent of it's actual effect on the outcome.

[10] For an excellent discussion of why the potential outcomes are independent of the observed exposure under exchangeability, see Chapter 2 of M. A. Hernán and Robins (Forthcoming)

Study Question 3: Why is the word "exchangeable" used to describe this concept? What, precisely, is being "exchanged"?

Although it seems that we have successfully written the potential risk as a function of the observed data, we are in need of two more assumptions. The first is **correct model specification**. This assumption is required when we rely on models to estimate effects, but can be minimized by using semi- or non-parametric approaches. There are several ways in which this assumption can be violated, and these include the omission of relevant interaction terms, or adjusting for continuous covariates using linear terms only. We will get into this issue in more depth when we discuss models.

Study Question 4: Can you think of a relation between correct model misspecification and exchangeability?

The second is **positivity**,[11] and requires exposed and unexposed individuals within all confounding levels (Mortimer et al. 2005,Westreich and Cole (2010)). There are two kinds of positivity violations

[11] Also known as the experimental treatment assignment assumption.

(non-positivity): structural (or deterministic) and stochastic[12] (or random). Structural non-positivity occurs when individuals with certain covariate values cannot be exposed. For example, in occupational epidemiology work-status (employed/unemployed in workplace under study) is a confounder, but individuals who leave the workplace can no longer be exposed to a work-based exposure. Alternatively, stochastic non-positivity arises when the sample size is not large enough to populate all confounder strata with observations. When faced with positivity violations, methods must be used that are less affected by positivity violations.[13] These include g estimation of a structural nested model, collaborative targeted minimum loss-based estimation, and the parametric g formula.

[12] The word **stochastic** is derived from the greek word "to aim," as in "to aim for a target."

[13] Warning: one cannot simply "avoid" positivity. In an extreme setting, non-positivity means that those who were exposed in the sample are very unlikely to be exposed (and vice versa). In such a situation, it may not make sense to estimate the average treatment effect, because there is a subset of the population who may never realistically be exposed (or unexposed). In this case, g estimation, cTMLE, and the parametric g formula can actually estimate parameters that differ slightly from the ATE.

Final Points

To summarize this section, let's review. We defined our causal effect of interest as the average treatment effect for a binary outcome / exposure, we we choose to be the causal risk difference:

$$E(Y^{x=1} - Y^{x=0})$$

where Y^x is the potential outcome that would be observed for a given individual if the exposure X were set to some value x. However, using observed data, we can only quantify the associational risk difference:

$$E(Y \mid X = 1) - E(Y \mid X = 0)$$

The causal risk difference is identifiable if we can re-write the associational risk difference as the causal risk difference, which we can do if we assume counterfactual consistency, no interference, exchangeability, positivity, and correct model specification.

However, these identifiability assumptions are specific to the average treatment effect. If we were intersted in a different estimand, we may need different assumptions. For example, if we wanted to estimate the local averate treatment effect:

$$E(Y^{x=1} - Y^{x=0} \mid X^{z=1} > X^{z=0})$$

where X is a binary indicator telling us if treatment were taken or

not, and Z is a binary indicator telling us if treatment were assigned or not. If the LATE is what we want to estiamte, then in addition to counterfactual consistency, no interference, exchangeability, positivity, and correct model specification, we also need the exclusion restriction:

$$Y^{xz} = Y^{x} \text{ for all } x \in \{0, 1\}$$

This assumption says that the treatment assignment indicator does not affect the outcome. We also need the "instrumentation" assumption:

$$E(X^{z=1} > X^{z=0}) \geq \delta > 0$$

References

Greenland, Sander. 2017. "For and Against Methodologies: Some Perspectives on Recent Causal and Statistical Inference Debates." *Eur J Epidemiol* 32 (1): 3–20. doi:10.1007/s10654-017-0230-6.

Greenland, Sander, and James Robins. 2009. "Identifiability, Exchangeability and Confounding Revisited." *Epidemiol Perspect Innov* 6 (1): 4.

Greenland, Sander, and JM Robins. 1986. "Identifiability, Exchangeability, and Epidemiological Confounding." *Int J Epidemiol* 15 (3): 413–19.

Greenland, Sander, James M. Robins, and Judea Pearl. 1999. "Confounding and Collapsibility in Causal Inference." *Stat Sci* 14 (1): 29–46.

Halloran, M Elizabeth, and Michael G Hudgens. 2016. "Dependent Happenings: A Recent Methodological Review." *Curr Epidemiol Rep* 3 (4): 297–305.

Hernan, M A, and S L Taubman. 2008. "Does Obesity Shorten Life? The Importance of Well-Defined Interventions to Answer Causal Questions." *Int J Obes* 32 (S3): S8–S14.

Hernán, M. A., and JM Robins. Forthcoming. *Causal Inference*. Boca Raton, FL: Chapman/Hall.

Hernán, Miguel A, and Tyler J VanderWeele. 2011. "Compound Treatments and Transportability of Causal Inference." *Epidemiol* 22 (3): 368–77. doi:10.1097/EDE.0b013e3182109296.

Hernán, Miguel A. 2005. "Invited Commentary: Hypothetical Interventions to Define Causal Effects—Afterthought or Prerequisite?" *Am J Epidemiol* 162 (7): 618–20.

Hudgens, M. G., and M. E. Halloran. 2008. "Toward Causal Inference with Interference." *J Am Stat Assoc* 103 (482): 832–42.

Imbens, Guido W, and Donald B Rubin. 2015. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. New York, NY: Cambridge University Press.

Krieger, Nancy, and George Davey Smith. 2016. "The Tale Wagged by the Dag: Broadening the Scope of Causal Inference and Explana-

tion for Epidemiology." *International Journal of Epidemiology* 45 (6): 1787–1808.

Mortimer, Kathleen M, Romain Neugebauer, Mark van der Laan, and Ira B Tager. 2005. "An Application of Model-Fitting Procedures for Marginal Structural Models." *Am J Epidemiol* 162 (4): 382–88. doi:10.1093/aje/kwi208.

Naimi, Ashley I., and Jay S. Kaufman. 2015. "Counterfactual Theory in Social Epidemiology: Reconciling Analysis and Action for the Social Determinants of Health." *Curr Epidemiol Reports* 2 (1): 52–60.

Naimi, Ashley I., Mireille E. Schnitzer, Erica E. M. Moodie, and Lisa M. Bodnar. 2016. "Mediation Analysis for Health Disparities Research." *American Journal of Epidemiology* 184 (4): 315–24. doi:10.1093/aje/kwv329.

Pearl, Judea, Madelyn R Glymour, and Nicholas Jewell. 2016. *Causal Inference in Statistics: A Primer*. United Kingdom: Wiley.

Robins, JM. 1987. "Addendum to 'a New Approach to Causal Inference in Mortality Studies with a Sustained Exposure Period–Application to Control of the Healthy Worker Survivor Effect'." *Comp Math Appl* 14 (9-12): 923–45.

Rubin, Donald B. 2005. "Causal Inference Using Potential Outcomes." *J Am Stat Assoc* 100 (469): 322–31.

Rubin, Donald B. 1980. "Randomization Analysis of Experimental Data: The Fisher Randomization Test Comment." *J Am Stat Assoc* 75 (371): 591–93.

Tchetgen Tchetgen, Eric J, and Tyler J VanderWeele. 2012. "On Causal Inference in the Presence of Interference." *Stat Methods in Med Res* 21 (1): 55–75.

Vandenbroucke, Jan P, Alex Broadbent, and Neil Pearce. 2016. "Causality and Causal Inference in Epidemiology: The Need for a Pluralistic Approach." *International Journal of Epidemiology* 45 (6): 1776–86.

VanderWeele, Tyler J, and Miguel Ángel Hernán. 2013. "Causal Inference Under Multiple Versions of Treatment." *Journal of Causal Inference* 1 (1): 1–20.

Westreich, Daniel, and Stephen R. Cole. 2010. "Invited Commen-

tary: Positivity in Practice." *Am J Epidemiol* 171 (6): 674–77.