Causal Inference for Time Dependent Treatments

Ashley I. Naimi, PhD

Outline

Causal Inference 1

- Correlation and Causation
- Introduction to Causal Inference
- Complex Longitudinal Data
- Potential Outcomes Notation
- Estimand, Estimator, Estimate

Correlation and Causation

In his *The Grammar of Science,* Karl Pearson (1911) wrote "[b]eyond such discarded fundamentals as 'matter' and 'force' lies still another fetish amidst the inscrutable arcana of modern science, namely, the category of cause and effect." He suggested that rather than pursue an understanding of cause-effect relations, scientists would be best served by measuring correlations through tables that classify individuals into specific categories. "Such a table is termed a contingency table, and the ultimate scientific statement of description of the relation between two things can always be thrown back upon such a contingency table."

Over a century later, a majority of statistics courses treat causal inference by simply stating that "correlation is not causation." This treatment is hardly sufficient, for at least two reasons: 1) As scientists, our primary interest is (should be) in cause-effect relations; 2) People continue to conflate correlation with causation[1]. For both of these reasons, we very much need to understand the conditions that would allow us to understand causality better. This is what "causal inference" is all about.

[1] Daniel Westreich and I reviewed a book in which the allure of "Big Data" was so strong, the authors quickly forgot that correlation $\neq$ causation. See Naimi and Westreich (2014)

I adopt the view that **the causal and statistical aspects of a scientific study should be kept as separate as possible.** The objective is to first articulate the conditions under which causal inference is possible, and then to understand what statistical tools will enable us to answer the causal question.[2] Causal inference tells us what we should estimate, and whether we can. Statistics tells us how to estimate it. By implication, we should avoid treating statistical models as if they were causal. Furthermore, to the best of our ability, we should avoid imposing unnecessary parametric assumptions on the causal models that we believe are generating the data. I will try to clarify what I intend by "imposing unnecessary parametric assumptions" later.

[2] Loosely speaking: Causal inference is the "what?" Statistics is the "how?" Epidemiology is the "why?"

Introduction to Causal Inference

"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.[3] The framework

[3] There are a number of introductory books and articles on causal inference in the empirical sciences. Here are some excellent options: Hernán and Robins (Forthcoming 2021), Pearl, Glymour, and Jewell (2016), Imbens and Rubin (2015)

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.[4] 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:

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

- 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?
- 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.

Complex Longitudinal Data

In this course, you have already encountered causal inference via potential outcomes when exposure under study is measured once (i.e., time fixed). In this lecture, we will focus on complex longitudinal data, and the complications that may arise when dealing with such 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.[5]

[5] 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

Scenario 2 is the classical situation that statisticians refer to as "longitudinal" data or correlated data. Here, we will deal with data that from scenarios 2 and 3. Repeated exposure, covariate, and/or outcome measurement is what leads to "longitudinal" data. But why complex?

Repeated measurement over time creates the opportunity for us to capture complex causal relations between past and future covariates. Suppose we measure an exposure 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," because we can use many standard methods we already know to analyze these data.

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

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.

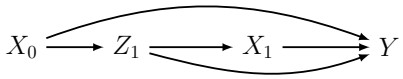

Figure 2: The simplest kind of complex longitudinal data. Note that the exposure at time zero affects the covariate at time 1 which affects the exposure at time 1. This feedback leads to confounding of the time 1 expsoure by a covariate that is affected by the prior exposure. Analysis of these data require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

Here, we will learn why this distinction is important, and how to use methods to account for this type of complex time-varying confounding.

Potential Outcomes Notation

The building blocks of modern causal inference are **potential outcomes** (Rubin 2005).

Importantly, these are conceptually distinct from **observed outcomes**. That is, the outcome that one might observe in a dataset is not the same as the potential outcome.

Potential outcomes are functions of exposures. For a given exposure x, we will write the potential outcome as Y^x.[6] **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

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

$Y^{x=1}$ (or, more succinctly, Y^0 or Y^1).

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

When the exposure 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 j, $\overline{X}_j$ denotes $\{X_0, X_1, X_2, \ldots X_j\}$. We can then define potential outcomes as a function of these exposure histories. For example, if the exposure was measured twice (once at time zero, and once at time 1), we could write $\overline{X}_1 = \{1, 1\}$, $Y^{\overline{x}_1=\overline{1}}$ is the outcome that would be observed if X_0 were set to 1 and X_1 were set to 1.

We could also generalize this notation to any number of follow-up times, such as for someone in a study with a follow-up time J, with a history indexed by $\{0, 1, 2, \ldots j, \ldots, J-2, J-1, J\}$. With the above notation, we could index the potential outcome that would be observed if someone were exposed to some value x up until some arbitrary time j as $Y^{\overline{X}_j=\overline{x}_j}$, or up until the end of their follow up time J as $Y^{\overline{X}_J=\overline{x}_J}$

Note that for an exposure variable measured numerous times over the course of follow-up, many potential outcomes can be defined. One need not restrict interest to contrasts under treatment strategies where everyone is exposed at all time points, versus where no one was exposed at all time points. In fact, for a binary treatment with up to J follow-up measurements, we can define up to 2^J treatment strategies (and thus, potential outcomes). The choice of which treatment strategy we select should be motivated by the research question of interest.

We can even define treatment strategies that depend on past covariates Z_j (Hernán and Robins Forthcoming 2021). These types of treatment strategies are common in HIV epidemiology, but can be found in other areas too. For example, in a recent study from our group ("The Effect of Preconception-Initiated Low-Dose Aspirin on Human Chorionic Gonadotropin–Detected Pregnancy, Pregnancy Loss, and Live Birth"), we looked at the effect of taking preconception low-dose aspirin on pregnancy outcomes among women at high risk of pregnancy loss. Among the treatment strategies we examined was the impact of taking aspirin starting at the 6th week of gestation. In this case, the potential outcome is defined based on a treatment strategy where women were asked to take placebo up until the 6th week of pregnancy, and then switched to LDA at the 6th week of pregnancy and thereafter. The treatment strategy is "dynamic" in the sense that it changes based on how other variables (in this case, hCG detected pregnancy status) change over follow-up.

Estimand, Estimator, Estimate

Causal inference starts with a clear idea of the effect of interest (the target causal parameter) (Petersen and Laan 2014). To do this, it helps to distinguish between estimands, estimators, and estimates.

STUDY QUESTION 2A: You are familiar 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.[7]

[7] 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 technically oriented,

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

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

There are many other estimands besides these.

All of the above estimands represent **average treatment effects** (on the risk difference, risk ratio, and odds ratio scale, respectively). This effect is also referred to as a marginal treatment effect, because it averages (or marginalizes) the effect over the entire sample. For instance, if we consider the risk ratio, it is easy to show that[8]

$$E(Y^1 - Y^0) = \sum_{i=1}^{N} Y_i^1 - \sum_{i=1}^{N} Y_i^0$$

[8] Recall that Y^x is not the observed (or sample) value of the outcome, so how do we actually get this average? When we discuss identifiability, we will see how we use observed data to quantify these contrasts.

However, we may want to estimate this effect in a subset of the population. For instance, $E(Y^1 - Y^0 \mid C = c)$ is the effect of $x = 1$ versus $x = 0$ among those with $C = c$. There are many different conditional treatment effects, this latter one being the simplest. Another common conditional treatment effect is the effect of treatment on the treated (ETT):

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

This effect compares the outcomes that would be observed if the exposure were set to 1 (Y^1) versus if the exposure were set to 0 (Y^0) among those who were observed to be exposed in the sample ($X = 1$).

To illustrate the relevance of this effect, consider the following (entirely fictional) scenario: During gestation of a high-risk pregnancy, two clinical options are available to manage the risk of death: induction of premature delivery induction and expectant management. Suppose a researcher is interested in quantifying the effect of inducing delivery prematurely on fetal and infant death. This researcher collects data on a cohort of high-risk pregnant women, including whether delivery was induced prematurely, fetal/infant death, and a host of confounding variables. All parties involved agree the study is designed perfectly (no confounding, measurement error, loss to follow-up). They calculate the average treatment effect of premature delivery induction on fetal and infant death on the risk difference scale:

$$E(Y^1 - Y^0) = 0.15$$

This researcher concludes that, if all high-risk pregnancies were induced prematurely ($X = 1$), 15 more out of every 100 would end in death, relative to what would happen if all high-risk pregnancies were left to expectant management ($X = 0$). In light of this incredibly high excess risk of death, this researcher would naturally advise abandoning the practice of premature delivery induction entirely.

Another researcher questions the relevance of the average treatment effect. They argue that physicians would never induce delivery prematurely in all versus no high-risk pregnancies. Rather, the more interesting question is: **for those women whose pregnancies were actually induced**, what would the risk of death have been had they not been induced? This researcher thus calculates the effect of treatment on the treated:

$$E(Y^1 - Y^0 \mid X = 1) = -0.05$$

This other researcher concludes that, among those whose pregnancies were actually delivered prematurely, the risk of death would have been higher had they not been delivered prematurely.

This example demonstrates the fundamental difference between the ATE and the ETT: for those high-risk pregnancies that were not induced prematurely, the act of inducing premature delivery would not be beneficial. But for those high-risk pregnancies that were induced prematurely, the act of inducing premature delivery was beneficial. The ATE averages the beneficial and non-beneficial effects in the entire population, to give an overall non-beneficial effect. The ETT isolates the beneficial effect among those who actually received the intervention. Thus, in this example, premature delivery actually did benefit those who received it, even though it would not benefit everybody.

There are many other estimands that can be defined, including the local average treatment effect, the survivor average causal effect, the complier average causal effect, and a host of principal strata effects. We will not discuss these in the context of this class, but be aware of their existence.

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 (e.g., 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, maximum likelihood, and the generalized method of moments as estimators:

```
# 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)

# the data
head(data.frame(CVD,smoking,confounder))
```

```
##   CVD smoking confounder
## 1   0       0          0
## 2   0       0          1
## 3   1       0          0
## 4   1       0          1
## 5   0       0          1
## 6   0       0          0
```

```
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
```

```
#GMM
round(gmm(CVD~smoking+confounder,x=cbind(smoking, confounder))$coefficients,4)
```

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

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

It is important to note that these are not causal risk differences, but are associational. To interpret them as causal effects, we have to evaluate whether we can identify the effect. We discuss this next.

Hernán, M. A., and JM Robins. Forthcoming 2021. *Causal Inference: What If*. Boca Raton, FL: Chapman/Hall.

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

Naimi, Ashley I., and Daniel J. Westreich. 2014. "Big Data: A Revolution That Will Transform How We Live, Work, and Think." *American Journal of Epidemiology* 179 (9): 1143–44.

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

Pearson, Karl. 1911. *The Grammar of Science*. 3rd ed. London, J.M. Dent & sons ltd.

Petersen, Maya L, and Mark J van der Laan. 2014. "Causal Models and Learning from Data: Integrating Causal Modeling and Statistical Estimation." *Epidemiol* 25 (3): 418–26. https://doi.org/10.1097/EDE.0000000000000078.

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.

"The Effect of Preconception-Initiated Low-Dose Aspirin on Human Chorionic Gonadotropin–Detected Pregnancy, Pregnancy Loss, and Live Birth." *Annals of Internal Medicine* 0 (0).