[SER 2021 Workshop] Causal Mediation: Modern Methods for Path Analysis

Iván Díaz, Nima Hejazi, Kara Rudolph

updated: March 23, 2021

Contents

List of Tables

List of Figures

Welcome to SER!

This open source, reproducible vignette is for a half-day workshop on modern methods for causal mediation analysis, given at the SER 2021 Meeting[1] on Monday, 24 May 2021.

[1]

About this workshop

Causal mediation analysis can provide a mechanistic understanding of how an exposure impacts an outcome, a central goal in epidemiology and health sciences. However, rapid methodologic developments coupled with few formal courses presents challenges to implementation. Beginning with an overview of classical direct and indirect effects, this workshop will present recent advances that overcome limitations of previous methods, allowing for: (i) continuous exposures, (ii) multiple, non-independent mediators, and (iii) effects identifiable in the presence of intermediate confounders affected by exposure. Emphasis will be placed on flexible, stochastic and interventional direct and indirect effects, highlighting how these may be applied to answer substantive epidemiological questions from real-world studies. Multiply robust, nonparametric estimators of these causal effects, and free and open source `R` packages (`medshift`[2] and `medoutcon`[3]) for their application, will be introduced.

To ensure translation to real-world data analysis, this workshop will incorporate hands-on `R` programming exercises to allow participants practice in implementing the statistical tools presented. It is recommended that participants have working knowledge of the basic notions of causal inference, including counterfactuals and identification (linking the causal effect to a parameter estimable from the observed data distribution). Familiarity with the `R` programming language is also recommended.

0.1 Workshop schedule

- 10:00A-10:30A: introductions/mediation set up
- 10:30A-11:00A: estimands and how to choose

[2] https://github.com/nhejazi/medshift

[3] https://github.com/nhejazi/medoutcon

- 11:00A-11:30A: discussion: how to choose in real-world examples
- 11:30A-12:00P: shift parameter introduction with application in lecture part
- 12:00P-12:15P break/discussion
- 12:15P-12:45P estimation for natural direct and indirect effects, interventional direct and indirect effects
- 12:45P-01:15P: practice `R` code for estimation
- 01:15P-01:30P: estimation for stochastic interventional direct and indirect effects
- 01:30P-01:50P: practice: code for estimation
- 01:50P-02:00P wrap up

NOTE: All times listed in Pacific Time.

About the instructors

Iván Díaz

My research focuses on the development of non-parametric statistical methods for causal inference from observational and randomized studies with complex datasets, using machine learning. This includes but is not limited to mediation analysis, methods for continuous exposures, longitudinal data including survival analysis, and efficiency guarantees with covariate adjustment in randomized trials. I am also interested in general semi-parametric theory, machine learning, and high-dimensional data.

Nima Hejazi

I am a PhD candidate in biostatistics at UC Berkeley, working under the joint direction of Mark van der Laan and Alan Hubbard. My research interests fall at the intersection of causal inference and machine learning, drawing on ideas from non/semi-parametric estimation in large, flexible statistical models to develop efficient and robust statistical procedures for evaluating complex target estimands in observational and randomized studies. Particular areas of current emphasis include causal mediation/path analysis, outcome-dependent sampling designs, targeted loss-based estimation, and applications in vaccine efficacy trials. I am also passionate about statistical computing and open source software development for applied statistics.

Kara Rudolph

I am an Assistant Professor of Epidemiology at Columbia University. My research interests are in developing and applying causal inference methods to understand social and contextual influences on mental health, substance use, and violence in disadvantaged, urban areas of the United States. My current work focuses on developing methods for transportability and mediation, and subsequently applying those methods to understand how aspects of the school and peer environments mediate relationships between neighborhood factors and adolescent drug use across populations. More generally, my work on generalizing/ transporting findings from study samples to target populations and identifying subpopulations most likely to benefit from interventions contributes to efforts to optimally target available policy and program resources.

0.2 Reproduciblity

These workshop materials were written using bookdown[4], and the complete source is available on GitHub[5]. This version of the book was built with R version 4.0.4 (2021-02-15), pandoc[6] version `r rmarkdown::pandoc_version()`, and the following packages:

package	version	source
bookdown	0.21.7	Github (rstudio/bookdown@b66380e)
bslib	0.2.4.9002	Github (rstudio/bslib@b2b4e55)
dagitty	0.3-1	CRAN (R 4.0.4)
data.table	1.14.0	CRAN (R 4.0.4)
downlit	0.2.1	CRAN (R 4.0.4)
dplyr	1.0.5	CRAN (R 4.0.4)
ggdag	0.2.3	CRAN (R 4.0.4)
ggfortify	0.4.11	CRAN (R 4.0.4)

[4] http://bookdown.org/

[5] https://github.com/tlverse/tlverse-handbook

[6] https://pandoc.org/

package	version	source
ggplot2	3.3.3	CRAN (R 4.0.4)
kableExtra	1.3.4	CRAN (R 4.0.4)
knitr	1.31	CRAN (R 4.0.4)
medoutcon	0.1.0	Github (nhejazi/medoutcon@f8f14c4)
medshift	0.1.4	Github (nhejazi/medshift@f9e11a9)
mvtnorm	1.1-1	CRAN (R 4.0.4)
origami	1.0.3	CRAN (R 4.0.4)
readr	1.4.0	CRAN (R 4.0.4)
rmarkdown	2.7.4	Github (rstudio/rmarkdown@1450461)
skimr	2.1.3	CRAN (R 4.0.4)
sl3	1.4.3	Github (tlverse/sl3@3950846)
stringr	1.4.0	CRAN (R 4.0.4)
tibble	3.1.0	CRAN (R 4.0.4)
tidyr	1.1.3	CRAN (R 4.0.4)

0.3 Setup instructions

0.3.1 R and RStudio

R and RStudio are separate downloads and installations. R is the underlying statistical computing environment. RStudio is a graphical integrated development environment (IDE) that makes using R much easier and more interactive. You need to install R before you install RStudio.

0.3.1.1 Windows

0.3.1.1.1 If you already have R and RStudio installed

- Open RStudio, and click on "Help" > "Check for updates". If a new version is available, quit RStudio, and download the latest version for RStudio.
- To check which version of R you are using, start RStudio and the first thing that appears in the console indicates the version of R you are running. Alternatively, you can type `sessionInfo()`, which will also display which version of R you are

running. Go on the CRAN website[7] and check whether a more recent version is available. If so, please download and install it. You can check here[8] for more information on how to remove old versions from your system if you wish to do so.

0.3.1.1.2 If you don't have R and RStudio installed

- Download R from the CRAN website[9].
- Run the `.exe` file that was just downloaded
- Go to the RStudio download page[10]
- Under Installers select RStudio x.yy.zzz - Windows XP/Vista/7/8 (where x, y, and z represent version numbers)
- Double click the file to install it
- Once it's installed, open RStudio to make sure it works and you don't get any error messages.

0.3.1.2 macOS / Mac OS X

0.3.1.2.1 If you already have R and RStudio installed

- Open RStudio, and click on "Help" > "Check for updates". If a new version is available, quit RStudio, and download the latest version for RStudio.
- To check the version of R you are using, start RStudio and the first thing that appears on the terminal indicates the version of R you are running. Alternatively, you can type `sessionInfo()`, which will also display which version of R you are running. Go on the CRAN website[11] and check whether a more recent version is available. If so, please download and install it.

0.3.1.2.2 If you don't have R and RStudio installed

- Download R from the CRAN website[12].

[7] https://cran.r-project.org/bin/windows/base/

[8] https://cran.r-project.org/bin/windows/base/rw-FAQ.html#How-do-I-UNinstall-R_003f

[9] http://cran.r-project.org/bin/windows/base/release.htm

[10] https://www.rstudio.com/products/rstudio/download/#download

[11] https://cran.r-project.org/bin/macosx/

[12] http://cran.r-project.org/bin/macosx

- Select the `.pkg` file for the latest R version
- Double click on the downloaded file to install R
- It is also a good idea to install XQuartz[13] (needed by some packages)
- Go to the RStudio download page[14]
- Under Installers select RStudio x.yy.zzz - Mac OS X 10.6+ (64-bit) (where x, y, and z represent version numbers)
- Double click the file to install RStudio
- Once it's installed, open RStudio to make sure it works and you don't get any error messages.

0.3.1.3 Linux

- Follow the instructions for your distribution from CRAN[15], they provide information to get the most recent version of R for common distributions. For most distributions, you could use your package manager (e.g., for Debian/Ubuntu run `sudo apt-get install r-base`, and for Fedora `sudo yum install R`), but we don't recommend this approach as the versions provided by this are usually out of date. In any case, make sure you have at least R 3.3.1.
- Go to the RStudio download page[16]
- Under Installers select the version that matches your distribution, and install it with your preferred method (e.g., with Debian/Ubuntu `sudo dpkg -i rstudio-x.yy.zzz-amd64.deb` at the terminal).
- Once it's installed, open RStudio to make sure it works and you don't get any error messages.

These setup instructions are adapted from those written for Data Carpentry: R for Data Analysis and Visualization of Ecological Data[17].

[13] https://www.xquartz.org/

[14] https://www.rstudio.com/products/rstudio/download/#download

[15] https://cloud.r-project.org/bin/linux

[16] https://www.rstudio.com/products/rstudio/download/#download

[17] http://www.datacarpentry.org/R-ecology-lesson/

Causal Mediation Analysis

[TO FILL IN]

The Roadmap for Statistical Learning

Learning Objectives

By the end of this chapter you will be able to:

1. Translate scientific questions to statistical questions.
2. Define a statistical model based on the knowledge of the experiment that generated the data.
3. Identify a causal parameter as a function of the observed data distribution.
4. Explain the following causal and statistical assumptions and their implications: i.i.d., consistency, interference, positivity, SUTVA.

Introduction

The roadmap of statistical learning is concerned with the translation from real-world data applications to a mathematical and statistical formulation of the relevant estimation problem. This involves data as a random variable having a probability distribution, scientific knowledge represented by a statistical model, a statistical target parameter representing an answer to the question of interest, and the notion of an estimator and sampling distribution of the estimator.

0.4 The Roadmap

Following the roadmap is a process of five stages.

1. Data as a random variable with a probability distribution, $O \sim P_0$.

2. The statistical model $\mathcal{M}$ such that $P_0 \in \mathcal{M}$.
3. The statistical target parameter Ψ and estimand $\Psi(P_0)$.
4. The estimator $\hat{\Psi}$ and estimate $\hat{\Psi}(P_n)$.
5. A measure of uncertainty for the estimate $\hat{\Psi}(P_n)$.

(1) Data: A random variable with a probability distribution, $O \sim P_0$

The data set we're confronted with is the result of an experiment and we can view the data as a random variable, O, because if we repeat the experiment we would have a different realization of this experiment. In particular, if we repeat the experiment many times we could learn the probability distribution, P_0, of our data. So, the observed data O with probability distribution P_0 are n independent identically distributed (i.i.d.) observations of the random variable $O; O_1, \ldots, O_n$. Note that while not all data are i.i.d., there are ways to handle non-i.i.d. data, such as establishing conditional independence, stratifying data to create sets of identically distributed data, etc. It is crucial that researchers be absolutely clear about what they actually know about the data-generating distribution for a given problem of interest. Unfortunately, communication between statisticians and researchers is often fraught with misinterpretation. The roadmap provides a mechanism by which to ensure clear communication between research and statistician – it truly helps with this communication!

The empirical probability measure, P_n

Once we have n of such i.i.d. observations we have an empirical probability measure, P_n. The empirical probability measure is an approximation of the true probability measure P_0, allowing us to learn from our data. For example, we can define the empirical probability measure of a set, A, to be the proportion of observations which end up in A. That is,

$$P_n(A) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}(O_i \in A)$$

In order to start learning something, we need to ask "What do we know about the probability distribution of the data?" This brings us to Step 2.

(2) The statistical model $\mathcal{M}$ such that $P_0 \in \mathcal{M}$

The statistical model $\mathcal{M}$ is defined by the question we asked at the end of Step 1. It is defined as the set of possible probability distributions for our observed data. Often $\mathcal{M}$ is very large (possibly infinite-dimensional), to reflect the fact that statistical knowledge is limited. In the case that $\mathcal{M}$ is infinite-dimensional, we deem this a nonparametric statistical model.

Alternatively, if the probability distribution of the data at hand is described by a finite number of parameters, then the statistical model is parametric. In this case, we subscribe to the belief that the random variable O being observed has, for example, a normal distribution with mean μ and variance σ^2. Formally, a parametric model may be defined

$$\mathcal{M} = \{P_\theta : \theta \in \mathcal{R}^d\}$$

Sadly, the assumption that the data-generating distribution has a specific, parametric form is all too common, especially since this is a leap of faith or an assumption made of convenience. This practice of oversimplification in the current culture of data analysis typically derails any attempt at trying to answer the scientific question at hand; alas, such statements as the ever-popular quip of Box that "All models are wrong but some are useful" encourage the data analyst to make arbitrary choices even when such a practice often forces starkly different answers to the same estimation problem. The Targeted Learning paradigm does not suffer from this bias since it defines the statistical model through a representation of the true data-generating distribution corresponding to the observed data.

Now, on to Step 3: "What are we trying to learn from the data?"

(3) The statistical target parameter Ψ and estimand $\Psi(P_0)$

The statistical target parameter, Ψ, is defined as a mapping from the statistical model, $\mathcal{M}$, to the parameter space (i.e., a real number) $\mathcal{R}$. That is, $\Psi : \mathcal{M} \to \mathbb{R}$. The estimand may be seen as a representation of the quantity that we wish to learn from the data, the answer to a well-specified (often causal) question of interest. In contrast to purely statistical estimands, causal estimands require identification from the observed data, based on causal models that include several untestable assumptions, described in more detail in the section on causal target parameters.

For a simple example, consider a data set which contains observations of a survival time on every subject, for which our question of interest is "What's the probability that someone lives longer than five years?" We have,

$$\Psi(P_0) = \mathbb{P}(O > 5)$$

This answer to this question is the estimand, $\Psi(P_0)$, which is the quantity we're trying to learn from the data. Once we have defined O, $\mathcal{M}$ and $\Psi(P_0)$ we have formally defined the statistical estimation problem.

(4) The estimator $\hat{\Psi}$ and estimate $\hat{\Psi}(P_n)$

To obtain a good approximation of the estimand, we need an estimator, an a priori-specified algorithm defined as a mapping from the set of possible empirical distributions, P_n, which live in a non-parametric statistical model, $\mathcal{M}_{NP}$ ($P_n \in \mathcal{M}_{NP}$), to the parameter space of the parameter of interest. That is, $\hat{\Psi} : \mathcal{M}_{NP} \to \mathbb{R}^d$. The estimator is a function that takes as input the observed data, a realization of P_n, and gives as output a value in the parameter space, which is the estimate, $\hat{\Psi}(P_n)$.

Where the estimator may be seen as an operator that maps the observed data and corresponding empirical distribution to a value in the parameter space, the numerical output that produced such a function is the estimate. Thus, it is an element of the parameter space based on the empirical probability distribution of the observed data. If we plug in a realization of P_n (based on a sample size n of the random variable O), we get back an estimate $\hat{\Psi}(P_n)$ of the true parameter value $\Psi(P_0)$.

In order to quantify the uncertainty in our estimate of the target parameter (i.e., to construct statistical inference), an understanding of the sampling distribution of our estimator will be necessary. This brings us to Step 5.

(5) A measure of uncertainty for the estimate $\hat{\Psi}(P_n)$

Since the estimator $\hat{\Psi}$ is a function of the empirical distribution P_n, the estimator itself is a random variable with a sampling distribution. So, if we repeat the experiment of drawing n observations we would every time end up with a different realization of our estimate and our estimator has a sampling distribution. The sampling distribution of some estimators can be theoretically validated to be approximately normally distributed by a Central Limit Theorem (CLT).

A class of Central Limit Theorems (CLTs) are statements regarding the convergence of the sampling distribution of an estimator to a normal distribution. In general, we will construct estimators whose limit sampling distributions may be shown to be approximately normal distributed as sample size increases. For large enough n we have,

$$\hat{\Psi}(P_n) \sim N\left(\Psi(P_0), \frac{\sigma^2}{n}\right),$$

permitting statistical inference. Now, we can proceed to quantify the uncertainty of our chosen estimator by construction of hypothesis tests and confidence intervals. For example, we may construct a confidence interval at level $(1-\alpha)$ for our estimand, $\Psi(P_0)$:

$$\hat{\Psi}(P_n) \pm z_{1-\frac{\alpha}{2}}\left(\frac{\sigma}{\sqrt{n}}\right),$$

where $z_{1-\frac{\alpha}{2}}$ is the $(1-\frac{\alpha}{2})^{\text{th}}$ quantile of the standard normal distribution. Often, we will be interested in constructing 95% confidence intervals, corresponding to mass $\alpha = 0.05$ in either tail of the limit distribution; thus, we will typically take $z_{1-\frac{\alpha}{2}} \approx 1.96$.

Note: we will typically have to estimate the standard error, $\frac{\sigma}{\sqrt{n}}$.

A 95% confidence interval means that if we were to take 100 different samples of size n and compute a 95% confidence interval for each sample, then approximately 95 of the 100 confidence intervals would contain the estimand, $\Psi(P_0)$. More practically, this means that there is a 95% probability that the confidence interval procedure generates intervals containing the true estimand value (or 95% confidence of "covering" the true value). That is, any single estimated confidence interval either will contain the true estimand or will not (also called "coverage").

0.5 Summary of the Roadmap

Data, O, is viewed as a random variable that has a probability distribution. We often have n units of independent identically distributed units with probability distribution P_0, such that $O_1, \ldots, O_n \sim P_0$. We have statistical knowledge about the experiment that generated this data. In other words, we make a statement that the true data distribution P_0 falls in a certain set called a statistical model, $\mathcal{M}$. Often these sets are very large because statistical knowledge is very limited - hence, these statistical models are often infinite dimensional models. Our statistical query is, "What are we trying to learn from the data?" denoted by the statistical target parameter, Ψ, which maps the P_0 into the estimand, $\Psi(P_0)$. At this point the statistical estimation problem is formally defined and now we will need statistical theory to guide us in the construction of estimators. There's a lot of statistical theory we will review in this course that, in particular, relies on the Central Limit Theorem, allowing us to come up with estimators that are approximately normally distributed and also allowing us to come with statistical inference (i.e., confidence intervals and hypothesis tests).

0.6 Causal Target Parameters

In many cases, we are interested in problems that ask questions regarding the effect of an intervention on a future outcome of interest. These questions can be represented as causal estimands.

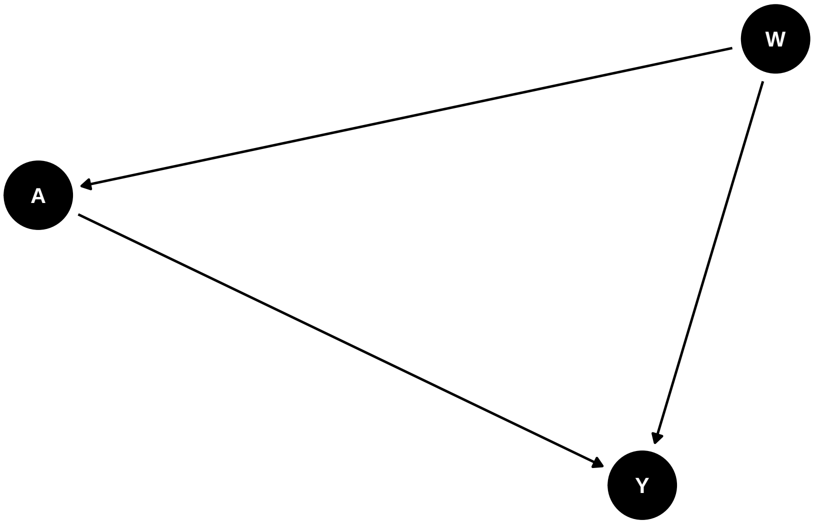
The Causal Model

After formalizing the data and the statistical model, we can define a causal model to express causal parameters of interest. Directed acyclic graphs (DAGs) are one useful tool to express what we know about the causal relations among variables. Ignoring exogenous U terms (explained below), we assume the following ordering of the variables in the observed data O. We do this below using `DAGitty` (Textor et al., 2011):

```r
library(dagitty)
library(ggdag)

# make DAG by specifying dependence structure
dag <- dagitty(
  "dag {
    W -> A
    W -> Y
    A -> Y
    W -> A -> Y
  }"
)
exposures(dag) <- c("A")
outcomes(dag) <- c("Y")
tidy_dag <- tidy_dagitty(dag)
```

```r
# visualize DAG
ggdag(tidy_dag) +
  theme_dag()
```



While directed acyclic graphs (DAGs) like above provide a convenient means by which to visualize causal relations between variables, the same causal relations among variables can be represented via a set of structural equations, which define the non-parametric structural equation model (NPSEM):

$$\begin{aligned} W &= f_W(U_W) \\ A &= f_A(W, U_A) \\ Y &= f_Y(W, A, U_Y), \end{aligned}$$

where U_W, U_A, and U_Y represent the unmeasured exogenous background characteristics that influence the value of each variable. In the NPSEM, f_W, f_A and f_Y denote that each variable (for W, A and Y, respectively) is a function of its parents and unmeasured background characteristics, but note that there is no imposition of any particular functional constraints(e.g., linear, logit-linear, only one interaction, etc.). For this reason, they are called non-parametric structural equation models (NPSEMs). The DAG and set of nonparametric structural equations represent exactly the same information and so may be used interchangeably.

The first hypothetical experiment we will consider is assigning exposure to the whole population and observing the outcome, and then assigning no exposure to the whole population and observing the outcome. On the nonparametric structural equations, this corresponds to a comparison of the outcome distribution in the population under two interventions:

1. A is set to 1 for all individuals, and
2. A is set to 0 for all individuals.

These interventions imply two new nonparametric structural equation models. For the case $A = 1$, we have

$$\begin{aligned} W &= f_W(U_W) \\ A &= 1 \\ Y(1) &= f_Y(W, 1, U_Y), \end{aligned}$$

and for the case $A = 0$,

$$\begin{aligned} W &= f_W(U_W) \\ A &= 0 \\ Y(0) &= f_Y(W, 0, U_Y). \end{aligned}$$

In these equations, A is no longer a function of W because we have intervened on the system, setting A deterministically to either of the values 1 or 0. The new symbols $Y(1)$ and $Y(0)$ indicate the outcome variable in our population if it were generated by the respective NPSEMs above; these are often called counterfactuals (since they run contrary-to-fact). The difference between the means of the outcome under these two interventions defines a parameter that is often called the "average treatment effect" (ATE), denoted

$$ATE = \mathbb{E}_X(Y(1) - Y(0)), \tag{1}$$

where $\mathbb{E}_X$ is the mean under the theoretical (unobserved) full data $X = (W, Y(1), Y(0))$.

Note, we can define much more complicated interventions on NPSEM's, such as interventions based upon rules (themselves based upon covariates), stochastic rules, etc. and each results in a different targeted parameter and entails different identifiability assumptions discussed below.

Identifiability

Because we can never observe both $Y(0)$ (the counterfactual outcome when $A = 0$) and $Y(1)$ (similarly, the counterfactual outcome when $A = 1$), we cannot estimate

1 directly. Instead, we have to make assumptions under which this quantity may be estimated from the observed data $O \sim P_0$ under the data-generating distribution P_0. Fortunately, given the causal model specified in the NPSEM above, we can, with a handful of untestable assumptions, estimate the ATE, even from observational data. These assumptions may be summarized as follows.

1. The causal graph implies $Y(a) \perp A$ for all $a \in \mathcal{A}$, which is the randomization assumption. In the case of observational data, the analogous assumption is strong ignorability or no unmeasured confounding $Y(a) \perp A \mid W$ for all $a \in \mathcal{A}$;
2. Although not represented in the causal graph, also required is the assumption of no interference between units, that is, the outcome for unit i Y_i is not affected by exposure for unit j A_j unless $i = j$;
3. Consistency of the treatment mechanism is also required, i.e., the outcome for unit i is $Y_i(a)$ whenever $A_i = a$, an assumption also known as "no other versions of treatment";
4. It is also necessary that all observed units, across strata defined by W, have a bounded (non-deterministic) probability of receiving treatment – that is, $0 < \mathbb{P}(A = a \mid W) < 1$ for all a and W). This assumption is referred to as positivity or overlap.

Remark: Together, (2) and (3), the assumptions of no interference and consistency, respectively, are jointly referred to as the stable unit treatment value assumption (SUTVA).

Given these assumptions, the ATE may be re-written as a function of P_0, specifically

$$ATE = \mathbb{E}_0(Y(1) - Y(0)) = \mathbb{E}_0 \left(\mathbb{E}_0[Y \mid A = 1, W] - \mathbb{E}_0[Y \mid A = 0, W]\right). \tag{2}$$

In words, the ATE is the difference in the predicted outcome values for each subject, under the contrast of treatment conditions ($A = 0$ vs. $A = 1$), in the population, averaged over all observations. Thus, a parameter of a theoretical "full" data distribution can be represented as an estimand of the observed data distribution. Significantly, there is nothing about the representation in 2 that requires parameteric assumptions; thus, the regressions on the right hand side may be estimated freely with machine learning. With different parameters, there will be potentially different identifiability assumptions and the resulting estimands can be functions of different components of P_0. We discuss several more complex estimands in later sections of this handbook.

The Natural Direct and Indirect Effects

[TO FILL IN]

The Interventional Direct and Indirect Effects

[TO FILL IN]

The Stochastic Direct and Indirect Effects

[TO FILL IN]

Bibliography

Textor, J., Hardt, J., and Knüppel, S. (2011). Dagitty: a graphical tool for analyzing causal diagrams. Epidemiology, 22(5):745.