APPLICATIONS OF AND TOOLS FOR CAUSAL INFERENCE

Bradley C. Saul

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Public Health in the Department of Biostatistics in the Gillings School of Global Public Health.

Chapel Hill
2017

Approved by:

Michael Hudgens

Steven Cole

Jess Edwards

John Preisser

Paul Stewart

Kihn Truong

ABSTRACT

Bradley C. Saul : APPLICATIONS OF AND TOOLS FOR CAUSAL INFERENCE
(Under the direction of Michael Hudgens)

Various topics related to causal inference with application to infectious disease and ecology are studied and software tools for such applications developed. The causal g-methods of Robins and colleagues – the parametric g-formula, marginal structural models, and structural nested models – are applied to a causal assessment of impaired water quality in North Carolina's Cape Fear River. The application demonstrates how a potential outcomes' causal analysis can be done with routine stream monitoring data. Under certain conditions, each of the g-methods can be cast in an estimating equation framework. Causal models often 'stack' estimating equations from multiple models, which can be a source of programming errors and bottlenecks. An R package for obtaining point and variance estimates from any arbitrary set of estimating equations is presented. The context of infectious diseases stimulated many advances in causal inference methods in the past 15 years. These methods and important contributions to the science of infectious diseases are reviewed.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: INTRODUCTION AND LITERATURE REVIEW

In this dissertation, I study causal inference applied to infectious disease and ecology and develop software for such applications. Chapter 2 causally assesses a significant water quality problem in North Carolina's Cape Fear River. The suite of g-methods developed by Robins (Robins, 1997; Robins and Hernán, 2009) are used to adjust for space- and time-varying confounding and demonstrate how causal effects of nutrient pollution can be estimated from stream surveillance data.

When parameterized appropriately, M-estimation (Stefanski and Boos, 2002) can be used to estimate causal effects from the g-methods. Once a set of estimating equations are defined, M-estimation theory provides the machinery for computing point estimates and asymptotically Normal variance-covariance estimates (Huber and Ronchetti, 2009). Chapter 3 presents an R package that translates the mathematical machinery of M-estimators into an application program interface (API) with which analysts can estimate parameters and covariance matrices from any set of estimating equations. The package also provides an API for finite sample variance corrections.

Theory and application of causal inference, especially in the realm of infectious disease studies, have received much attention in the past 10-15 years. Chapter 4 reviews causal inference methods applicable to studies of infectious diseases and reflects on how the infectious disease context stimulated many advances in causal inference. The review includes discussion of interference, principle stratification, test-negative study design, causal surrogates, and other topics.

The next sections provide background for each of these topics, after first introducing causal inference.

1.1 Causal inference versus statistical inference

Statistical inference, broadly speaking, is about quantifying properties of a population's distribution from a sample. Scientists often want to understand more than mere associations within a population's joint probability distribution. The scientist wants to know how a distribution *changes* when acted upon: does X *cause* Y? Understanding causal relationships requires more than just examining a probability distribution. Causal inference cannot be done by the statistician alone with his bag of probability tricks. Pearl (2010a) clarifies my point: "An associational concept is any relationship that can be defined in terms of a joint distribution of observed variables, and a causal concept is any relationship that cannot be defined from the distribution alone." Additional non-data assumptions must be made. Fortunately, such assumptions can be notated mathematically. As Rubin (2005) notes, "... notation explicitly representing ... potential outcomes is an exceptional contribution to causal inference."

1.1.1 Notation of potential outcomes

Early in the history of modern statistics, Neyman introduced the concept of potential outcomes (Splawa-Neyman et al., 1990), applying the idea to yields under different varieties of seed spread on agricultural plots. A potential outcome is simply the outcome a unit of the population would have had if, possibly contrary to fact, it had been subject to a certain exposure in a particular way. Rubin (1978) is often attributed for propelling the use of potential outcomes in the late 20th century, but many have contributed to advancing potential outcomes.

Throughout, I use the following generic notation. Let the random variable A_i be an exposure or intervention of interest for unit $i = 1, \dots n$ in a study, and a_i be fixed, realized values of the exposure. I use the terms intervention, treatment, and exposure interchangeably. Observed (random) outcomes are denoted Y_i and potential outcomes denoted $Y_i(a)$. Other measured covariates are labeled L_i and unmeasured covariates U_i. I will generally drop the i index notation except when it may not be clear. Boldface indicates vectors or matrices as appropriate. For example, $\mathbf{A} = \{A_1, \ldots, A_n\}$.

1.1.2 Primacy of the causal estimand

> "Before anything else, the right question needs to be framed." (Freedman, 2005)

Causal inference demands careful attention to the target of inference. More so than in traditional statistical literature, causal inference emphasizes the estimand. Rubin (2005) even labels the estimand as "The Science". Entire papers are written on causal estimands and their identifiability assumptions (e.g., Robins and Greenland, 1992; Ogburn and VanderWeele, 2014). Rubin (2005) urges practitioners to establish their estimand before concerning themselves with statistical technicalities.

An estimand, in Rubin's formulation, includes all of the units' potential outcomes. But unit-level effects are generally not identified due to the *fundamental problem of causal inference* (Holland, 1986). The fundamental problem says that an individual may have many potential outcomes each depending on a particular treatment assignment. But in a given study, only one vector of treatment is assigned, hence the fundamental problem that only one potential outcome can be observed for each subject. The causal consistency theorem (Pearl, 2010b) links observed outcomes to potential outcomes: when A is set to a, the potential outcome equals the observed outcome, $Y(a) = Y$.

The fundamental problem is not completely stifling. Population-level summaries of unit-level potential outcomes can often be identified and estimated. Indeed, the population-level summary is the part of the estimand generally of most interest. But specifications of unit-level potential outcomes are just as important, particularly when relaxing the standard stable unit treatment value assumption (SUTVA).

SUTVA (Rubin, 1986) restricts an individual's set of potential outcomes in order to facilitate identification. SUTVA bundles two assumptions: no hidden forms of treatment and no interference. The assumption of no hidden forms of treatment (well-defined treatments or treatment variation irrelevance (VanderWeele, 2009)) is best illustrated with a medical study example. The assumption says that no matter how a subject takes a medication, the treatment effect on a given individual is the same. If taking the drug with a meal or without a meal caused two different effects, the assumption would be violated.

1.1.3 Assumption of no interference

SUTVA's second assumption of no interference has received significant attention in recent years (Hudgens and Halloran, 2008; Sobel, 2006). No interference limits the number of potential outcomes to those determined only by the unit's possible treatment settings. The treatment of other individuals cannot affect the outcome of individual i; i.e., $Y(\mathbf{a}) = Y(a_i)$. For a binary treatment, this means a subject has two potential outcomes — one under treatment, one under control. For many research problems, and the applications in this dissertation, this assumption is too strict. Nutrient pollution at an upstream location in a stream is likely to affect stream productivity at downstream locations. Vaccination of some individuals against a highly contagious disease is likely to affect whether others will become infected.

1.1.4 Identifiability assumptions

Identification of an estimand relies on two primary assumptions. First, potential outcomes must be independent of the intervention, possibly conditional on covariates: $Y(a) \perp A|L = l$. This assumption goes by many names: no unmeasured confounding (NUC), ignorability, and exchangeability. The strong version claims ignorability for all levels of a, whereas the weak version claims ignorability for certain levels of a. For most estimation methods, the strong version is needed for identifiability, but there are cases where the weak version suffices (Vansteelandt and Joffe, 2014). The second assumption, known as positivity or the experimental treatment value assumption, says that each level of the intervention has some non-zero probability of occurring for all levels of the confounding covariates: $\Pr(A = a|L = l) > 0$.

Together, exchangeability and positivity form the basis for most identifiability problems in causal inference. In terms of positivity, if some level of exposure structurally has zero probability of occurrence, then clearly we have zero chance of learning how that level of exposure affects the outcome. Some applied problems run into practical positivity problems. While positivity holds in the population, the observed probability for a certain level may be zero or near zero in a particular sample.

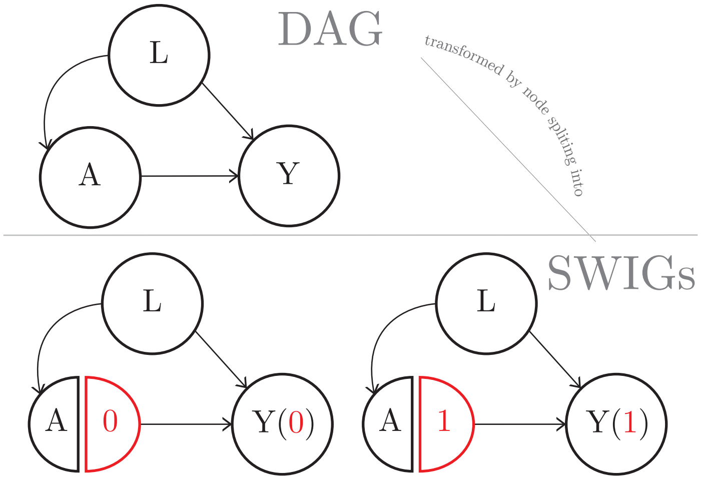
A randomized trial elucidates exchangeability. When treatment is randomized, a subject's potential outcomes are independent of the treatment assignment. The subject's potential outcomes did not depend on the treatment assignment, thus treatment assignment is *ignorable*. When the randomized assignment works correctly, the treated and control groups are similar in every aspect *except* receiving the treatment.

The above are the essential theoretical assumptions necessary to claim causality from association, given data for the entire population. Generally we don't have data on the entire population and work from sample data. Often, due to the curse of dimensionality, parametric or semiparametric models are components of a causal model. In order for conditional exchangeability to hold, the component models must be correctly specified. In practice, particularly in observational studies, the component models are almost surely never specified precisely, but the hope is that the models are reasonably accurate.

1.1.5 Causal graphs

Directed acyclic graphs, or DAGs, are a well-studied and useful tool for reasoning about confounding (Pearl, 2010a; Joffe et al., 2012). A longstanding barrier for translating between the notation of DAGs and potential outcomes is that DAGs do not encode potential outcomes. Richardson and Robins (2013) introduced the Single World Intervention Graph (SWIG) as a tool to unite the graphical and potential outcomes approaches.

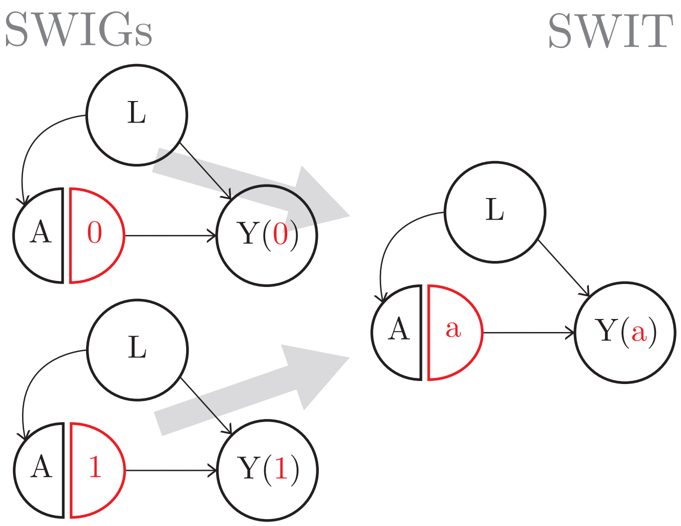
Figure 1.1: A directed acyclic graph (DAG) is transformed into single world intervention graphs (SWIGs) by splitting intervention nodes. Each SWIG represents one level of a. In this case, the SWIGs are the same for both levels.



As seen in Figure 1.1, the obvious distinction between a DAG and a SWIG is the split node. The node-splitting operation of the SWIG splits the intervention nodes into the random part and the fixed (intervened upon) part. This action makes conditional independence relations between the intervention and the potential outcome more easily seen.

Throughout this dissertation, I use SWIGs to represent causal problems. More specifically, I use Single World Intervention Templates (SWITs). A SWIG technically represents the graph for a single world; that is, a unique level of the intervention. With a SWIT, the same graph is assumed to hold for all levels of the intervention that the SWIT represents (Figure 1.2).

Figure 1.2: When the SWIGs for all levels of a are isomorphic, a single world intervention template (SWIT) summarizes SWIGs into a single graph.



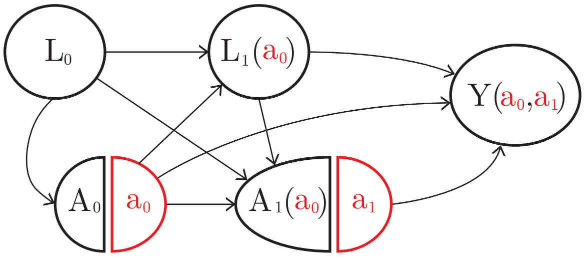
1.2 Motivating problems

1.2.1 Algae on the Cape Fear

North Carolina's Cape Fear watershed extends from the source of the Deep and Haw rivers near Greensboro to the river's mouth in Wilmington. Beginning in 2009, toxic algal blooms, sometimes reaching up to 70 miles in length, occurred in the middle portion of the river – downstream from Fayetteville into the coastal plain (Isaacs et al., 2014; Cape Fear River Partnership, 2013). These blooms impacted water quality and threatened fisheries during the summer months of 2009 to 2012. While low flows and high temperatures during those years certainly contributed to igniting algal production, it is unknown how nutrient pollution specifically contributed to the blooms and which upstream regions of the river contributed most substantially.

When measuring causal effects from multiple upstream locations simultaneously, space-varying confounding may be present. For example, consider three locations, ordered from upstream to downstream. The exposure at site 1 could be confounded by a covariate at

Figure 1.3: Time-varying exposures with two time points shows three properties: L_1 (1) is associated with the potential outcome, (2) is affected by A_0, and (3) affects $A_1(a_0)$.



site 1 which in turn affects both the exposure at site 2 as well as the outcome at site 3. This upstream to downstream structure, though played out over space, is analogous to time-varying confounding. Figure 1.3 illustrates this type of confounding.

Since his seminal paper in 1986, James Robins along with several colleagues developed the g-methods for estimating causal effects in the presence of time-varying confounding: the G-computation algorithm of the g-formula, inverse probability weighting of marginal structural models, and G-estimation of structural nested models. These methods may prove valuable for ecological causal assessments as in the Cape Fear River nutrient pollution problem.

1.2.2 g-methods

The g-methods were designed to consistently estimate joint exposure effects in the presence of time-varying confounding. All three methods have slightly different assumptions and different estimation techniques. The methods are described briefly (and perhaps loosely) here, but are further developed in Chapter 2.

The g-formula is a general decomposition of counterfactual outcomes, $E[Y(\overline{a})] = \int_{\overline{l}} E[Y|\overline{A} = \overline{a}, \overline{L} = \overline{l}]dF_{\overline{l}}$, where $\overline{V}_t$ denotes the history of a generic random variable V up to time t. Estimating the right hand side of the g-formula may be challenging. Generally, simple parametric models (e.g., linear or logistic regression) are fit for $E[Y|\overline{A} = \overline{a}, \overline{L} = \overline{l}]$ and $dF_{\overline{l}}$. When $\overline{L}$ is multivariate, this may mean fitting many models. The G-computation algorithm uses these models to simulate pseudo-populations from which counterfactual values can be estimated.

Marginal structural models (MSMs) posit a model for the counterfactual mean as a function of the exposure history, $E[Y(\overline{a})] = h(\beta; \overline{a}) = (e.g.)\beta_0 + \beta_1 \sum_{t=1}^{T} a_t$. To adjust for confounding, weights are used to create a pseudo-population in which confounding is absent.

An MSM may include baseline effect modifiers in its parameterization, but the effect of a is presumed homogeneous across levels of $\overline{L}$. Structural Nested Models (SNMs) do not suffer this limitation, which, according to Vansteelandt and Joffe (2014) are a "partially realized promise."

Unlike the g-formula or MSMs, SNMs directly models a causal contrast – generally that of removing treatment: $E[Y(\overline{a}) - Y(\overline{0})|\overline{L}] = \psi(\overline{A}, \overline{L}; \beta)$. Software is not readily available for SNMs, and parameterizing the models requires careful attention. In many cases, SNMs may be more challenging to fit, requiring a p-dimensional grid search where p is the number of causal parameters. But in some cases, SNMs can be cast in an estimating equation framework and estimation can proceed in the M-estimation framework.

1.2.3 Programming causal estimators is a leading cause of headaches among graduate students

Causal models, especially those based on the g-methods, are often composed of one or more component models. As an example, a logistic regression model with parameters α may be used to estimate propensity scores, $\Pr(A|L; \alpha)$. The fitted propensity scores $\hat{p} = \Pr(A|L; \hat{\alpha})$ are then a component in the causal estimators. The α parameters are solutions to the score equations, $\sum_i \psi(O_i; \alpha) = 0$, where O_i represents the data for unit i. In many cases, a target causal estimand, say μ, may be the solution to $\sum_i \psi(O_i; \mu) = \sum_i CE_i(\hat{p}) - \mu = 0$, where CE stands for some generic causal estimator of μ. When parameters can be written as solutions to a set of estimating equations, M-estimation theory (Huber and Ronchetti, 2009; Stefanski and Boos, 2002) provides the mathematics which will obtain consistent and asymptotically Normal estimators under weak regularity conditions. The popular method of generalized estimating equations (GEE) (Liang and Zeger, 1986) for fitting marginal models is a variant of M-estimation. But as Stefanski and Boos (2002) discuss in their primer, M-estimation is more broadly applicable.

In the example of the previous paragraph, the estimating equation functions for both sets of parameters can be 'stacked' so that estimates can be simultaneously obtained for all parameters. That is, $\psi(O_i;\theta) = (\psi_{i,\alpha}(O_i;\alpha), \psi_{i,\mu}(O_i;\mu))^\intercal$ can be used to obtain point and variance estimates for the entire set of parameters, $\theta = \{\alpha, \mu\}$. Often estimators for the component model parameters are common regression estimators, and off-the-shelf statistical software can be used. The target parameter estimator in many cases has a closed form, making programming of the estimator relatively straightforward.

M-estimation of the variance-covariance matrix has the form $A_m^{-1} B_m \{A_m^{-1}\}^T$, where $A_m = \sum_{i=1}^m A_i$ and $B_m = \sum_{i=1}^m B_i$, and so only requires two quantities from the estimating equations, $A_i = (\partial\psi(O_i;\theta)/\partial\theta)$ and $B_i = \psi(O_i;\theta)\psi(O_i;\theta)^T$. An estimate of B_i is simply the outer product of the estimating equations with $\hat{\theta}$ plugged in. A_i is the matrix of partial derivatives of the ψ functions with respect to θ. In simple cases, these derivatives may be expressed analytically, but often many elements of A_i do not have a closed form, thus requiring numerical approximations.

In my experience, and in observing others program variance estimators, variance M-estimators are challenging to translate into software. With multiple component models, there are many places to make programming errors. Robins et al. (2000) suggest practitioners can avoid this issue by treating the parameters from the component models as known. This approach is unnecessarily conservative as it is well known that estimating the component parameters in the variance estimation improves standard error estimates.

Many software packages have programmed specific cases of M-estimators. R packages such as `geepack` (Halekoh et al., 2006), `gee` (Carey, 2015), and `geeM` (McDaniel et al.,

2013) implement GEE, and others solve variants of GEE methods. The package `sandwich` (Zeileis, 2006) obtains the M-estimation (a.k.a "sandwich") variance estimates from various model types, but only from one model at a time. To date, no package offers a generic, flexible framework for M-estimation.

As a functional programming language, R is suited for abstracting mathematical specification into software. In R, functions can return functions (Wickham, 2014). I propose to use this simple concept for an M-estimation package. The user provides a function that takes in data, maps the relationship between data and parameters, and returns a function that takes in parameters. This last function is then used for the computations of M-estimation: finding the root(s) of the estimating equations and Jacobian matrix (A_i) of the equations evaluated at the root(s). In this way, the poor graduate student only needs to program one function, $\psi(O_i; \theta)$, in order to do M-estimation.

1.2.4 Causal inference in infectious disease

Infectious diseases complicate textbook causal inference. For example, basic causal inference assumes an individual's potential outcomes do not depend on the intervention status of other individuals. But infectious diseases, by definition, depend on interaction among hosts, resulting in such effects as "herd immunity." This greatly increases the number of potential outcomes an individual may have. With a binary intervention, a subject could have as many potential outcomes as possible treatment assignments in the study, or 2^n.

Dating back to the origins of modern statistics in agricultural experiments, many recognized between-plot interference as a nuisance. In the causal world, Rubin's single unit treatment value assumption (SUTVA) (Rubin, 1980) essentially nullifies interference effects. In infectious disease, interference is not a nuisance, but a process of interest (Halloran and Struchiner, 1995).

The problem of interference in infectious diseases has motivated much recent causal inference research. In chapter 4, I will review how the context of infectious diseases stimulated many advances in causal inference, beyond just interference. I will also study how the science of infectious diseases has benefited from a rigorous causal inference framework. The chapter will include time-varying confounding, test negative studies, negative controls, regression discontinuity, principal surrogates, interference, and contagion effects.

1.3 Outline

In Chapter 2, I use the causal g-methods developed by Robins and colleagues to assess causal effects of a significant water quality impairment in North Carolina's Cape Fear River. The methods – g-formula, marginal structural models, and structural nested models – are modified to handle the space and time interference structure of stream ecosystems. Results from these models will be compared.

In Chapter 3, I introduce an R software package for estimating parameters and covariances from a set of estimating equations. While several packages exist that implement specific uses of estimating equations, in particular the GLM-like Generalized Estimating Equations of Liang and Zeger (1986) (`gee`, `geepack`, etc), none allow the user to define a custom set of estimating equations. My package `geex` will allow the analyst to estimate parameters and covariance from any set of estimating equations they can dream up from standard M-estimation theory (Stefanski and Boos, 2002).

In Chapter 4, I review and discuss methods of causal inference related to infectious disease interventions.

CHAPTER 2: UPSTREAM CAUSES OF DOWNSTREAM EFFECTS

2.1 Introduction

Nutrient pollution of U.S. streams costs billions of dollars each year (Dodds et al., 2009). The EPA calls reducing nutrient pollution in U.S. waterways a “high priority” (EPA, 2015) and acknowledges that Nitrogen-Phosphorous (NP) pollution is a causal factor in algal blooms. However, the EPA’s 2015 report also notes that since many factors may contribute to a harmful algal bloom (HAB), “it is often difficult or impossible to say *how much more* likely an HAB is because of nutrient pollution.” Lack of experimental manipulation and small sample sizes are among many potential pitfalls in making causal inferences using stream surveillance data (Norton et al., 2014). The EPA’s Causal Analysis/Diagnosis Decision Information System (CADDIS) outlines a reasoned, methodical process for assessing causality in stream ecosystems (Norton et al., 2009). Suter et al. (2002), among the primary developers of CADDIS, state that data analysis methods in causal assessments “should be selected to best illuminate the association given the amounts and types of data available.” In this chapter, a potential outcome (or counterfactual) approach is considered for drawing inference about the causal effects of nutrient pollution on stream ecosystems.

Data on North Carolina’s Cape Fear River is analyzed as a case study. This is a large Piedmont-Coastal Plain system that is representative of many riverine systems from Virginia south through North Florida (Dame et al., 2000). Formerly considered a moderately productive river (Kennedy and Whalen, 2007), in 2009 it began experiencing harmful algal blooms consisting of the cyanobacterium (blue-green alga) *Microcystis aeruginosa* near Lock and Dam 1 (LD1) that reappeared periodically through 2012 (Isaacs et al., 2014). Freshwater algal blooms are often stimulated by phosphorus (P) loading (Howarth and

Marino, 2006), but in Coastal Plain rivers and streams, algal blooms are largely stimulated by nitrogen (N) loading (Mallin et al., 2004; Dubbs and Whalen, 2008). In North Carolina (NCDENR, 2005) as well as many states and provinces, regulatory agencies regularly monitor concentration of the algal pigment chlorophyll *a* as a proxy for algal bloom strength. Long-term monitoring of this river by state-certified coalitions, including the Lower Cape Fear River Program and Middle Cape Fear Coalition, has provided a data set of nutrients, chlorophyll *a*, and other water quality parameters for the middle and lower river, where the blooms are concentrated.

This chapter shows that causal effects of upstream nutrient concentrations on downstream chlorophyll *a* can be estimated from observational water quality data. Correlation analyses or regression techniques, while invaluable for exploring associations within an ecosystem, do not typically estimate causal effects. With publicly available watershed monitoring data, we assess causal effects of nutrient concentrations measured upstream of LD1 on chlorophyll *a* levels at LD1 by adapting the causal g-methods (Robins and Hernán, 2009; Hernán and Robins, 2017). Originally developed for assessing the effect of a time-varying exposure, here the g-methods are extended to the setting where exposure varies in both time and space. In particular, the causal models allow for spatial interference (Verbitsky-Savitz and Raudenbush, 2012; Di Gennaro and Pellegrini, 2016) in the sense that exposure (nutrient concentration) at one location may affect the outcome (chlorophyll *a*) at another location. Inference about parameters of marginal structural models, the parametric g-formula, and structural nested models which accommodate the spacetime interference structure of a stream ecosystem is considered using estimating equation theory (Stefanski and Boos, 2002), with small sample adjustments (Fay and Graubard, 2001) to account for limited independent replicates.

The chapter is organized as follows. Section 2.2 motivates the analysis and describes the available data on the Cape Fear River. Section 2.3 introduces potential outcomes, key assumptions, and the target estimand. A graphical representation of the model assumptions using a Single World Intervention Template (Richardson and Robins, 2013) is also presented. The g-methods are presented in Section 2.4 along with small sample variance corrections. The simulation study in Section 2.5 validates and compares statistical properties of the g-methods. The Cape Fear River data are analyzed in Section 2.6. We discuss our findings and their limitations in Section 2.7. Sections 2.8, 2.9, 2.10, and 2.11 contain additional mathematical derivations and stability analyses.

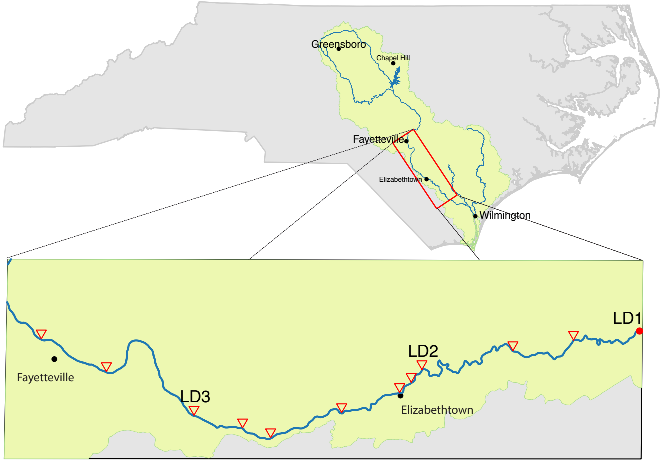
2.2 Motivation, materials, and notation

2.2.1 Cape Fear River nutrient pollution and algal blooms

During the summers of 2009-2012, algal blooms unprecedented in scale and composition occurred near LD1 near Kelly, NC. Isaacs et al. (2014) reported that samples collected from these blooms in 2009 and 2012 consisted predominantly of toxic *Microcystis aeruginosa* cyanobacteria. The multi-stakeholder watershed action plan for the Cape Fear River identifies blue-green algae, *M. aeruginosa* in particular, as a significant threat to the river ecosystem (Cape Fear River Partnership, 2013). Over 2 million people rely on drinking water from the Cape Fear watershed, and algal blooms have impacted taste and odor from some water treatment plants (Ahuja, 2013). Brunswick County, in southeastern North Carolina, obtains some of its drinking water directly from the river near LD1. Taste and odor problems arising from the cyanobacterial blooms forced the water utility to increase its level of water treatment, at significant cost, to produce acceptable drinking water. Thus, causes of the recent degradation in Cape Fear River water quality are key management concerns.

The 9000 square mile Cape Fear watershed is contained entirely within the political boundaries of North Carolina, extends from Greensboro to Wilmington, and includes parts of Durham and Chapel Hill. The Cape Fear River forms at the confluence of the Haw and Deep Rivers and once supported rich fisheries of anadromous fish (Cape Fear River Partnership, 2013). Figure 2.4 shows the extent of the Cape Fear watershed and the area of interest for this study, the section of river from Fayetteville to LD1.

Figure 2.4: The map shows the extent of the Cape Fear watershed within the political boundaries of North Carolina, as well as the region of interest for this study. The algal blooms generally occurred near LD1. This study examines causal relationship between nutrient concentration measured at upstream sampling locations (open triangles) on chlorophyll *a* at LD1.

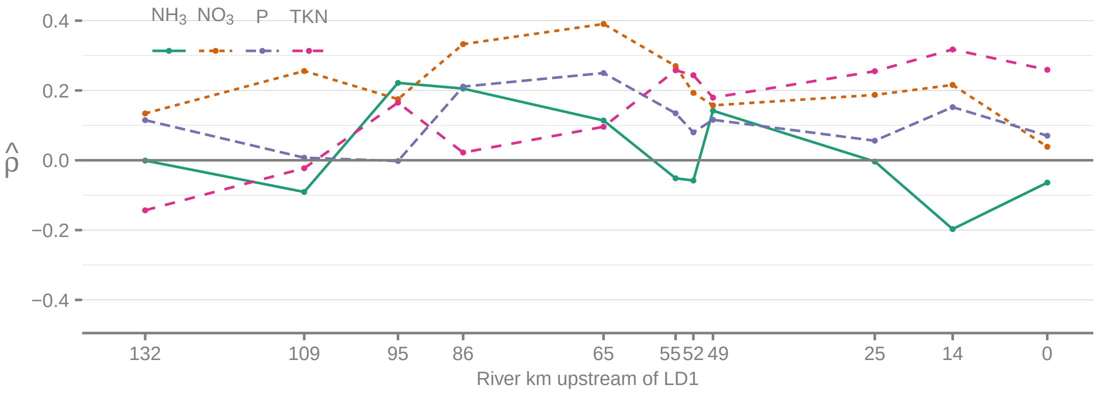


The Nature Conservancy of North Carolina obtained coalition-produced, state-certified data consisting of monthly measurements from locations throughout the Cape Fear basin from July 1996 through June 2013. Prior to 1999, chlorophyll *a* was not consistently measured at LD1. Since large blooms at LD1 were reported mainly during summer months, we focused our analysis on observations from June, July, August, and September of 1999 to 2012 from the main stem of the Cape Fear River upstream of LD1. The data include concentration measurements of four NP compounds (all in mg/L): nitrate (NO_3), ammonia (NH_3), total Kjeldahl nitrogen (TKN), and phosphorous (P).

2.2.2 Associations of nutrients and LD1 chlorophyll

A simple correlation analysis shows generally positive associations between upstream nutrients and chlorophyll *a* levels at LD1. Figure 2.5 plots Spearman's correlation coefficients between nutrient concentrations at sampling locations within the study region and LD1 chlorophyll *a*. Each nutrient has a slightly different trajectory over the course of

Figure 2.5: Spearman's correlation coefficients between nutrient levels at upstream sampling locations and LD1 chlorophyll a, using observations from June-September of 1999-2012.



the river, but with the exception of TKN, the correlation peaks between 65 and 95 river kilometers upstream of LD1. These associations suggest a relationship between upstream nutrient levels and LD1 chlorophyll. Our goal in this chapter is to adjust for confounding to determine to what degree the upstream nutrients cause changes in LD1 chlorophyll a.

2.2.3 A mathematical description

Let $i = 1, \ldots, m$ index independent replicates; for the Cape Fear data, $m = 14$ corresponding to the years 1999 to 2012. We assume that clusters of summer months are sufficiently far enough apart in time to be considered independent. That is, observations from June to September of year i are independent of the same set of observations in year $i' \neq i$, but observations within a year may be correlated. Let W be a generic variable indexed as W_{ist}, where i indicates year, $s = 1, \ldots, S$ indicates the sampling locations, ordered from upstream to downstream, and $t = 1, \ldots, T$ indicates the month (e.g., $t = 1 =$ June). In the sequel, the i notation is dropped where convenient.

Observed values of chlorophyll a (μg/L) are $\log_2$ transformed and denoted as Y_{st}. To avoid confusion with notation, chlorophyll a is referred to as chlorophyll in the sequel. The effect of each nutrient is considered separately in our analysis, and nutrient exposure is generically denoted as A_{st}. Other covariates measured concurrently with nutrient concentrations include the date and time of measurements, temperature (°C), dissolved oxygen, pH, and turbidity. In addition to covariates recorded in the water quality data set, daily mean discharge data from stream gauges located at the William O'Huske Lock and Dam (LD3) and LD1 were downloaded from USGS and converted to m^3/s. Discharge values were linearly interpolated based on river distance for sampling locations between the gauges. For each location, the average of the mean daily discharge from the same date as the water quality measurements plus the two prior days was used in the analysis. Let L_{kst} denote the kth ($k = 1, \ldots, p$) covariate measured at location s in month t, where the vector L_{st} of length p collects all covariates for a particular location and time. Let $O_{st} = \{Y_{st}, A_{st}, L_{st}\}$ denote the observed random variables at location s in month t. For any variable W_{st}, let the $s \times t$ matrix $\overline{W}_{st}$ denote the variable's history for all locations upstream to and including location s, plus all time points prior to and including time point t.

2.3 Causal inference from upstream to downstream

Let $Y_{s^\star t}(\overline{a}_{st})$ be the potential value of $\log_2$ chlorophyll at location $s^\star$ in month t had the exposure history been $\overline{a}_{st}$, for $s < s^\star$. By causal consistency (Pearl, 2010a), $Y_{s^\star t}(\overline{a}_{st}) = Y_{s^\star t}$ when $\overline{A}_{st} = \overline{a}_{st}$. Define the average potential outcomes for a location of interest $s^\star$ over months $t = 1, \ldots, T$ as $E\{[Y_{s^\star 1}(\overline{a}_{s1}), Y_{s^\star 2}(\overline{a}_{s2}), \ldots, Y_{s^\star T}(\overline{a}_{sT})]^\intercal\} = E[Y_{s^\star}(\overline{a}_s)]$. In the analysis, $s^\star = 3$ corresponds with LD1.

2.3.1 Effects of interest

Policymakers' or scientists' effects of interest, or estimands, can be stated in terms of functions of average potential outcomes. For example, what difference would be expected, on average, in LD1 chlorophyll levels if NH$_3$ exposure at the upstream points LD3 and LD2 was set to be above, rather than below, a certain threshold during the month of June ($t = 1$)? Letting s_1 correspond to LD3 and s_2 correspond to LD2, this estimand is, in the notation defined above, $E[Y_{31}((a_{11}, a_{21})^\intercal) - Y_{31}((a'_{11}, a'_{21})^\intercal)]$, where a_{11} and a_{21} indicate

NH_3 exposure at LD3 and LD2 above the NH_3 threshold and a'_{11} and a'_{21} indicate NH_3 levels below the threshold.

Consider the estimand which measures the effect of setting nutrient concentrations at two upstream locations, s_1 and s_2, on LD1 chlorophyll averaged across T months, i.e.,

$$\mu = \frac{1}{T} \sum_{t=1}^{T} E\left[Y_{3t}([0_{t-1}, a_t]) - Y_{3t}(0_t)\right], \tag{2.1}$$

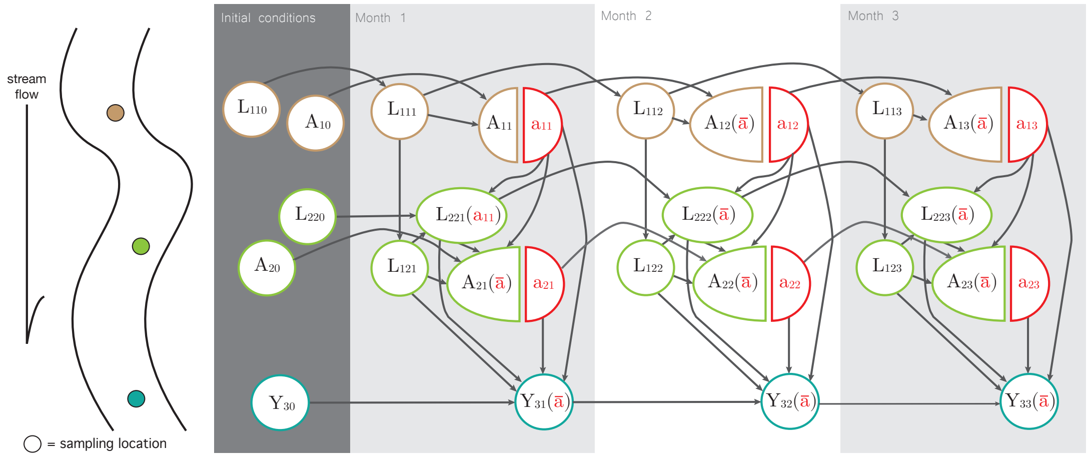
where 0_t is a $2 \times t$ matrix of zeros defined as the empty set when $t = 0$, $a_t = (a_{1t}, a_{2t})^{\intercal}$, and $[U, V]$ indicates the concatenation of matrices U and V. As written, μ is defined for any exposure setting a. In the Cape Fear River analysis, exposure is defined as a binary variable being above ($a = 1$) or below ($a = 0$) cutpoints specified in Section 2.6.

The estimand (2.1) characterizes the average effect on LD1 chlorophyll when intervening at two upstream locations simultaneously. This parameter is of interest to the community of scientists working on the Cape Fear River who want to understand the effects of nutrient concentrations from different upstream locations on LD1 chlorophyll during the summer when the toxic algal blooms generally occurred. Including exposures from two upstream locations allows demonstration of how to handle covariates that affect the exposure and vary between upstream locations as well as the potential for bias when such covariates are not accounted for correctly.

2.3.2 Single world intervention graph

Richardson and Robins (2013) introduced single world intervention graphs (SWIGs) to unify the graphical approach to causal inference (e.g., see Pearl, 2009) and the more algebraic potential outcomes framework (e.g., see Rubin, 2005). An important difference between the approaches is the representation of potential outcomes. Algebraic notation can easily distinguish between potential and observed outcomes ($Y(a)$ versus Y). Directed acyclic graphs (DAGs) do not explicitly encode potential outcomes. SWIGs do.

Figure 2.6: This single world intervention template conceptualizes the upstream to downstream process. A_{st} is an exposure of interest, and a_{st} are fixed values of the exposure. L_{22t} are space-varying confounding covariates which (i) are associated with the potential outcome, $Y_{3t}(\overline{a})$, (ii) affect A_{2t}, and (iii) are affected by A_{1t}.



Reading a SWIG is similar to reading a DAG. Nodes represent variables and edges suggest causal relationships between nodes (Figure 2.6). In a SWIG, however, intervention nodes are transformed by a node-splitting operation. Instead of a single A_{11} node as in a DAG, the A_{11} semicircle represents the random variable for exposure at location 1 at time 1. The a_{11} semicircle represents the fixed setting of the exposure (possibly contrary to fact) at the same spacetime point. Figure 2.6 is technically a single world intervention template (SWIT), not a SWIG. SWITs are a graphical template for a set of exposure levels, whereas SWIGs represent the graph for a single exposure level. We assume the SWIGs have the same form for all exposures and all their levels, hence a single SWIT describes the SWIGs for all exposure levels.

2.4 Estimation of causal effects

It is well known that time-varying confounding may introduce bias if not accounted for in estimation. Robins and Hernán (2009) describe three g-methods for estimating counterfactuals from observational data in the presence of time-varying confounding: the parametric g-formula, fitting marginal structural models (MSMs) using inverse probability weighting, and g-estimation of structural nested models (SNMs). This section describes extensions of these g-methods to the spacetime setting.

2.4.1 Causal assumptions

The causal effect μ can be identified by the distribution of the observable random variables by considering the structure of a stream (represented by Figure 2.6) as a sequentially and conditionally randomized experiment. Given (i) covariate values up to and including location s and month t and (ii) values of the past exposure(s) prior to location s and month t, A_{st} is assumed to be independent of the potential outcomes. The covariates in $\overline{L}_{st}$ must block all back-door paths between A_{st} and $Y_{s^\star t}(\overline{a}_{st})$ (Pearl, 2009), which implies conditional independence, commonly referred to as the strong ignorability or no unmeasured confounding assumption:

$$Y_{s^\star t}(\overline{a}_{st}) \perp A_{st} | \overline{L}_{st}, \overline{A}^{\circ}_{st}, \tag{2.A1}$$

where $\overline{A}^{\circ}_{st} = \overline{A}_{st} \setminus \{A_{st}\}$.

Let $f_w = f(w)$ be the probability density or mass function for a random variable W. For the g-formula and MSMs, identification of causal effects also depends on a positivity assumption $f(a_{st}|\overline{O}^{\circ}_{st} = \overline{o}^{\circ}_{st}) > 0$ for all $\overline{o}^{\circ}_{st}$ such that $f(\overline{O}^{\circ}_{st}) > 0$ where $\overline{O}^{\circ}_{st} = \overline{O}_{st} \setminus \{A_{st}\}$. That is, each level of exposure must have some non-zero probability of occurring at all spacetime points for all possible covariate and exposure histories.

These assumptions are needed to identify causal effects nonparametrically. In many applications, as in ours, common finite-dimensional parametric models such as linear or logistic regression are employed to model aspects of the distribution of observable random variables. These models must be correctly specified in order for the resulting inferences to be valid.

2.4.2 Parametric g-formula

The g-formula is a mathematical identity which relates the distribution of counterfactuals to the distribution of the observable random variables (Robins and Hernán, 2009). For example, using the g-formula, the counterfactual mean can be expressed as:

$$E[Y_{s^\star t}(\overline{a}_{st})] = \int_{\overline{l}_{st}} E[Y_{s^\star t}|\overline{A}_{st} = \overline{a}_{st}, \overline{L}_{st} = \overline{l}_{st}] f_{\overline{l}_{st}}. \tag{2.2}$$

where $f_{\overline{l}_{st}} = \prod_{j=1}^{s} \prod_{k=1}^{t} f_{l_{jk}|\overline{l}_{j-1,k-1}, \overline{a}_{j-1,k-1}}$. In practice, the mean model $E[Y_{s^\star t}|\overline{A}_{st} = \overline{a}_{st}, \overline{L}_{st} = \overline{l}_{st}]$ and conditional densities $f_{l_{jk}|\overline{l}_{j-1,k-1}, \overline{a}_{j-1,k-1}}$ are not known, and estimated values $\hat{E}[Y_{s^\star t}|\overline{A}_{st} = \overline{a}_{st}, \overline{L}_{st} = \overline{l}_{st}]$ and $\hat{f}_{\overline{l}_{st}}$ are plugged into (2.2) to estimate $E[Y_{s^\star t}(\overline{a}_{st})]$. Though these quantities may be estimated nonparametrically for a single spacetime point, a parametric approach may be necessary to estimate more complicated quantities such as μ. In both the analysis and simulations presented below, the mean model was parameterized as a linear model with main effects only for A_{2t}, A_{1t}, and L_{2t}, with corresponding parameters β_1^g, β_2^g, and β_4^g, respectively. Let γ_3^g be the parameter corresponding to A_{1t} in a simple linear model for the density of L_{2t}. When the exposure settings are binary where $a_{st} = 1$ and $a'_{st} = 0$ for all t, then

$$\mu = \frac{1}{T} \sum_{t=1}^{T} \{\beta_1^g + (\beta_2^g + \beta_4^g \gamma_3^g)\} = \beta_1^g + \beta_2^g + \beta_4^g \gamma_3^g.$$

Section 2.9 contains the algebraic details. Maximum likelihood is used to estimate the model coefficients. The coefficient estimates are then plugged into (2.2) to obtain the estimator of $E[Y_{s^\star t}(\overline{a}_{st})]$.

2.4.3 Marginal structural model

Marginal structural models posit a parametric relationship between an exposure history and a counterfactual outcome. Consider the following MSM:

$$E[Y_{s^\star t}(\overline{a}_{st})] = \beta^m_{0t} + \beta^m_1 a_{st} + \beta^m_2 a_{s-1,t}. \tag{2.3}$$

Each month may have a distinct intercept β^m_{0t}, but the counterfactual mean depends only on exposure at two upstream locations during the same month. From (2.3), $\mu = \beta^m_1 + \beta^m_2$. Parameters in MSMs can be estimated consistently using inverse probability weighting methods (Hernán et al., 2000). We use the stabilized inverse probability weight where each observed outcome is weighted by:

$$SW_{st} = \prod_{j=1}^{s}\prod_{k=1}^{t} \frac{f(a_{jk}|\overline{A}^\circ_{jk} = \overline{a}^\circ_{jk})}{f(a_{jk}|\overline{O}^\circ_{jk} = \overline{o}^\circ_{jk})}. \tag{2.4}$$

The product is taken across the dimensions of space s and time t as opposed to a single dimension as in Robins et al. (2000). Logistic regression is used to estimate $f(a_{st}|\overline{A}^\circ_{st} = \overline{a}^\circ_{st})$ and $f(a_{st}|\overline{O}^\circ_{st} = \overline{o}^\circ_{st})$. Weighting observed outcomes by (2.4), generalized estimating equations (GEE) (Liang and Zeger, 1986) with an independence working correlation matrix are used to estimate β^m.

2.4.4 Structural nested (mean) model

Instead of modeling counterfactual means from which causal contrasts are then derived, structural nested models directly model a causal effect. In general, SNMs model the effect of removing treatment within strata l, $E[Y(a) - Y(0)|L = l]$. Vansteelandt and Joffe (2014) describe several advantages of SNMs over MSMs. For one, target parameters in SNMs are identified with weak ignorability rather than strong ignorability and without the positivity assumption. Strong ignorability assumes (2.A1) holds for all $\overline{a}_{st}$, whereas weak ignorability assumes (2.A1) holds only for $\overline{a}_{st} = \overline{0}_{st}$. Also, the asymptotic variance of IPW estimators tend to be highly sensitive to misspecification of the exposure model(s), to which G-estimators of SNM parameters tend to be less sensitive (Vansteelandt and Joffe, 2014).

The Cape Fear River analysis uses the following structural nested mean model:

$$E\begin{bmatrix} \vdots \\ Y_{3t}\begin{pmatrix} & a_{1t} \\ 0_t & \\ & a_{2t}\end{pmatrix} - Y_{3t}\begin{pmatrix} & a_{1t} \\ 0_t & \\ & 0\end{pmatrix} |\overline{L}_{2t} = \bar{l}_{2t} \\ Y_{3t}\begin{pmatrix} & a_{1t} \\ 0_t & \\ & 0\end{pmatrix} - Y_{3t}\begin{pmatrix} & 0 \\ 0_t & \\ & 0\end{pmatrix} |\overline{L}_{1t} = \bar{l}_{1t} \\ \vdots \end{bmatrix} = \begin{pmatrix} \vdots \\ \psi_{1t}(\overline{a}_t, \bar{l}_{2t}; \beta^s) \\ \psi_{2t}(\overline{a}_t, \bar{l}_{1t}; \beta^s) \\ \vdots \end{pmatrix} = \begin{pmatrix} \vdots \\ \beta_1^s a_{2t} \\ \beta_2^s a_{1t} \\ \vdots \end{pmatrix}. \tag{2.5}$$

For $T = 4$, (2.5) has dimension 8×1, as each t has two ψ functions for each month. The ψ function corresponds to a "blip down" process (Vansteelandt and Joffe, 2014), removing the effect of treatment one spatial location at a time. The first ψ_{1t} "blips out" and quantifies the effect of a_{2t}, and ψ_{2t} quantifies the effect of a_{1t}. This SNM assumes that ψ does not depend on l; that is, the causal contrast does not include interactions between exposure and covariates. Thus, $\beta_1^s + \beta_2^s = \mu$.

To estimate β^s, a vector $U_t(\beta^s)$ is constructed whose mean value, given $\overline{L}_{st}$ and $\overline{A}_{st}$, equals the mean outcome that would have been observed had treatment been removed:

$$U_t(\beta^s) = \begin{pmatrix} U_{1t}(\beta^s) \\ U_{2t}(\beta^s) \end{pmatrix} = \begin{pmatrix} Y_{3t} - \beta_1^s A_{2t} \\ Y_{3t} - (\beta_1^s A_{2t} + \beta_2^s A_{2,t-1}) \end{pmatrix}.$$

Using a modified version of equation (33) in Vansteelandt and Joffe (2014), the solution to the following estimating equations is a consistent estimator for β^s:

$$\begin{aligned} &\sum_i \sum_{s=1}^{S} \sum_{t=1}^{T} \left[d_s(\overline{L}_{ist}, \overline{A}_{ist}) - E\{d_s(\overline{L}_{ist}, \overline{A}_{ist}) | \overline{L}_{ist}, \overline{A}_{ist}\} \right] \\ &\qquad\qquad \left[U_{ist}(\beta^s) - E\{U_{ist}(\beta^s) | \overline{L}_{ist}, \overline{A}_{ist}\} \right] = 0 \end{aligned} \tag{2.6}$$

where d is some arbitrary distance function. Per the suggestion of Vansteelandt and Joffe (2014), we let $d_s = E\left\{\partial U_{st}(\beta)/\partial\beta|\overline{L}_{st}, \overline{A}_{st}\right\}$. The solution to (2.6) is called a G-estimator and has the advantage of double robustness. That is, the estimator is consistent when either the model for the transformed outcome $E[U_{st}(\beta)|\overline{L}_{st}, \overline{A}_{st}]$ or the exposure model (which is a component of $E[d_s(\overline{L}_{st}, \overline{A}_{st})|\overline{L}_{st}, \overline{A}_{st}])$ are correctly specified. Parametric regression models were used to model both the outcome and exposure (linear and logistic regression, respectively). In some cases, solving (2.6) yields a closed form solution for $\hat{\beta}^s$ as shown in Section 2.10.

2.4.5 Estimating equation inference

In each of the previous three sections, the g-formula (2.2), MSM (2.3), and SNM (2.5) were specified such that parameter estimates may be obtained by solving a set of unbiased estimating equations. Therefore, under certain regularity conditions, the estimators will be consistent and asymptotically normal, and the empirical sandwich variance estimator can be used to consistently estimate the asymptotic covariance matrix of the model parameter estimators. In the case of the g-formula, the target estimand μ is a function of β_1^g, β_2^g, β_4^g, and γ_3^g, so the estimated variance of $\hat{\mu}$ can be obtained using the delta method. For MSMs, observed outcomes for each time point are weighted by estimated values of (2.4), and weighted generalized estimating equations are used to obtain $\hat{\beta}^m$. Variance estimates for MSMs can be obtained by stacking the score equations for the parametric models used to estimate the weights plus the estimating equations corresponding to (2.3) weighted by (2.4). Point estimates in the SNM were obtained from the closed form of $\hat{\beta}^s$, while variance estimates were obtained by stacking the score equations of both the outcome and exposure models along with estimating equation (2.6).

For all three methods, consistent variance estimators follow from estimating equations (i.e., M-estimation) theory (Stefanski and Boos, 2002). Let $\hat{\theta}$ be estimators that are solutions to a set of p equations $\sum_{i=1}^{m}\psi(O_i, \hat{\theta}) = 0$, where ψ is a vector of functions of length p corresponding to the number of parameters in θ. From our causals models, θ contains the target parameters β plus any nuisance parameters present in estimating the IP weights, outcome model, or exposure model. The asymptotic covariance for $\hat{\theta}$ is $\Sigma = A^{-1}B\{A^{-1}\}^{\intercal}/m$, where $A_i = \partial\psi(O_i, \theta)/\partial\theta$, $A = E[A_i]$, $B_i = \psi(O_i, \theta)\psi(O_i, \theta)^{\intercal}$, and $B = E[B_i]$. The empirical sandwich variance estimator replaces the expectations with their empirical counterparts and θ with $\hat{\theta}$; e.g., $\hat{A} = m^{-1}\sum_{i=1}^{m} \partial\psi(O_i, \theta)/\partial\theta|_{\theta=\hat{\theta}}$.

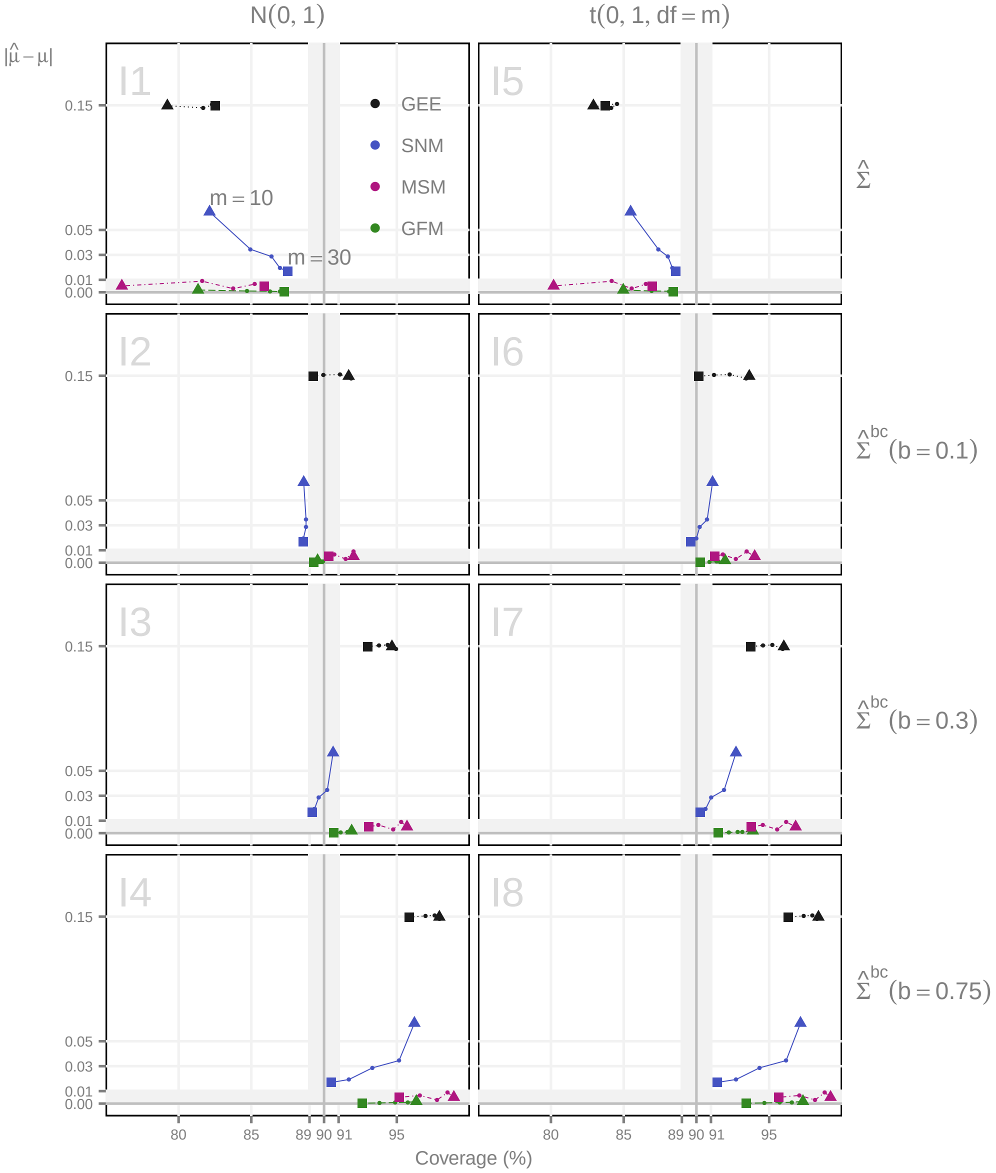
The empirical sandwich variance estimator is asymptotically consistent but tends to underestimate the true variance in small samples (Fay and Graubard, 2001; Li and Redden, 2015). We examine the bias corrected estimator of Fay and Graubard (2001) in simulations. The bias corrected variance estimator replaces B_i with $B_i^{bc}(b) = H_i(b)B_iH_i(b)^{\intercal}$ to form $\Sigma^{bc} = A^{-1}B^{bc}(b)\{A^{-1}\}^{\intercal}$, where $H_i(b) = \{1 - \min(b, \{A_iA\}_{jj})\}^{-1/2}$ and $\{A_iA\}_{jj}$ denotes the jjth element of A_iA. The constant b is less than 1 and chosen by the analyst intended to prevent extreme corrections when A_iA is close to 1. Fay and Graubard (2001) “arbitrarily” set $b = 0.75$. To explore the sensitivity of our variance estimates we used $b = 0.1, 0.3$, and 0.75. Variance estimates were used to construct Wald confidence intervals based on either a normal distribution or a *t* distribution with m degrees of freedom.

2.5 Simulation study

Based on the SWIT in Figure 2.6, we used the simcausal R package (Sofrygin et al., 2016) to simulate data for $m = 10, 15, 20, 25,$ and 30 years. For each m, 24,000 data sets were generated according to the parametrization provided in Section 2.8. For each simulated data set, estimates and 90% confidence intervals of μ were computed using the causal g-methods, plus a naive GEE approach that ignores space- and time-varying confounding. Code for the simulations is available upon request.

Figure 2.7 shows the absolute bias of the four estimators and empirical coverage of the corresponding confidence intervals for each set of simulations. The horizontal shading highlights bias < 0.01, and the vertical shading highlights coverage within 1% of the nominal 90% coverage. Each facet in the figure shows a different correction to the variance estimator and distribution used in constructing the Wald confidence intervals. For all three causal methods, the absolute bias shrinks and the coverage improves as sample size (i.e., number of years) increases. The bias for the MSM and g-formula (GFM) methods is smaller compared to the SNM for all m analyzed in Figure 2.7. The SNM still has an average absolute bias of about 0.01 even when $m = 30$. In a secondary simulation of 1000 data sets where $m = 150$ (not shown), the SNM bias shrunk to the same order of magnitude as the MSM and g-formula methods. The naive GEE estimator is always biased, empirically illustrating why methods that account for time- and/or space-varying confounding should be used when such effects are present. For small m, the unadjusted variance estimator performs poorly, covering in the 75 - 85% range for all the methods. As m increases, the unadjusted coverage based on a normal distribution improves. The bias correction tends to overcorrect and have coverages greater than 90% for all settings of b.

Figure 2.7: Simulation results show absolute bias $|\hat{\mu} - \mu|$ on the y-axis. Each line shows results for a method from $m = 10$ (triangle) to $m = 30$ (square). The proportion of simulations where the 90% confidence interval included the true value is shown on the x-axis. Each box shows results for the different methods of forming confidence intervals, with the columns defining the distribution and the row defining the form of the variance estimator. The bias is unaffected by corrections to the confidence intervals, hence the y-coordinates do not change across the boxes.



2.6 Cape Fear River analysis

Beginning with the sampling location 132 km upstream near I-95 in Fayetteville, we estimate μ for a given NP species at that site ($s = 2$) and the site just upstream ($s = 1$). That is, the causal effect was estimated for setting an exposure at location A and location B, then the effect of location B and location C, then the effect of location C and D, and so on downstream until the last sampling location upstream of LD1.

The methods described above treat exposures as binary, but the species of NP are measured on a continuous scale. For each species and set of upstream locations, the exposures were discretized using three cutpoints (1st, 2nd and 3rd quartiles) at the two upstream locations during April to October of 1999. The distribution for each NP species varied over the course of the river, so a single river-wide cutpoint would not be meaningful. At some locations, for example, all the observed concentrations could be below a river-wide cutpoint. For each nutrient, five space-varying confounding covariates were considered: each of the other three nutrients, dissolved oxygen, or pH.

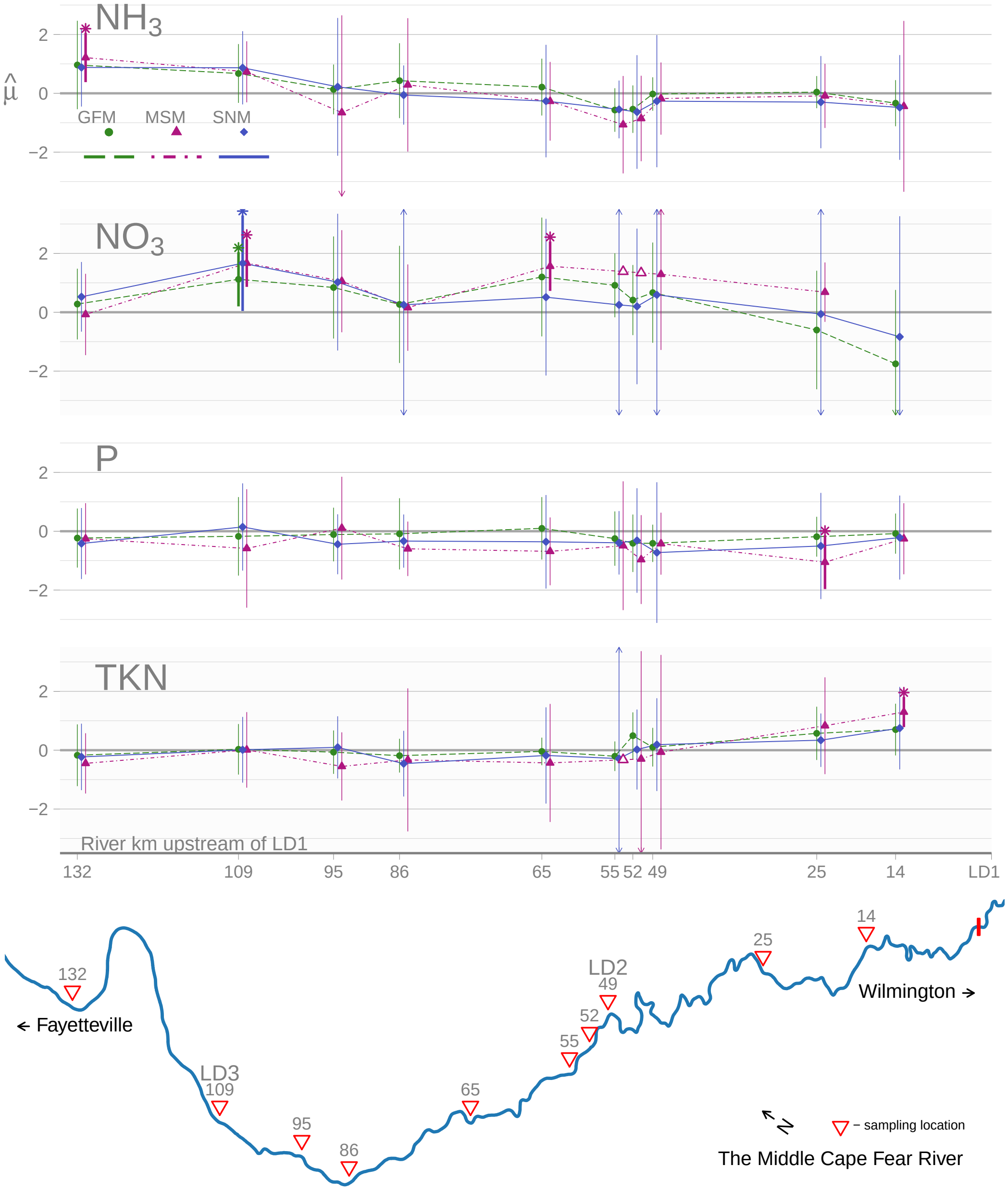
In summary, 600 causal effects were estimated (4 NP species × 10 sets of s_2 and s_1 upstream locations × 3 exposure cutpoints × 5 space-varying confounding covariates) using each of the four different analysis methods (g-formula, MSM, SNM, and GEE). Wald confidence intervals and p-values were computed using the eight combinations of distribution and variance estimator as in Figure 2.7. For each analysis method, the outcome and exposure models were parameterized similarly to the simulations, with the exception that L_{1st} was not a single covariate and instead both temperature and discharge were included in the models. As a sensitivity analysis, causal effects were also estimated using different parameterizations for the outcome and exposure models.

For the observations considered in our analysis, at most 5% of the data were missing for any given covariate. LD1 chlorophyll values were missing for May 2000 and July 2004. These two missing values were replaced by taking the average from the month prior and post.

Of the 770 observations for the 11 upstream locations during May to September of 1999 to 2012, 5 NH_3, 5 NO_3, 22 TKN, 38 P, and 16 temperature values were missing. No NH_3 or NO_3 were missing for June to September. Where possible, missing covariate values at a location-month were singly imputed in the following sequential manner. First, we attempted to average the values immediately upstream and immediately downstream during the same month. If either value immediately upstream or immediately downstream was missing, then we averaged values from the next site upstream and next site downstream during the same month. If neither of these approaches imputed a value, then we averaged the prior month and next month from the same location. This approach imputed missing values for all nutrient covariates except P, which had missing values for all summer months at all of our upstream locations in 2009. We excluded 2009 in analyses involving P.

Figure 2.8 shows causal effects with point-wise 90% confidence intervals for the four nutrient species. This figure shows results using the median cutpoint (as described above). GEE and g-formula results were largely similar, so we only show g-formula results. The space-varying covariate is P for the nitrogen species, and it is NH_3 for P. In this set of analyses, a non-target model failed to converge when estimating the MSM in four cases (three for NO_3 and one TKN). These are indicated by open points, which were interpolated from the other values, unless the model failed for the location just upstream of LD1. Confidence intervals are based on the bias corrected standard errors ($b = 0.3$) and a t-distribution with m degrees of freedom. Thick bars with stars above indicate statistically significant estimates at the $\alpha = 0.1$ level.

Figure 2.8: Results of causal analysis of Cape Fear River data from June-September of 1999-2012 for each of the four measured nutrients. Points are estimates of μ. Vertical lines are 90% confidence intervals with thicker lines highlighting intervals that do not cross zero. The intervals used in this plot use method I7 of Figure 2.7. Open points are where the model fitting algorithm failed to converge.



The point estimates of the three causal methods tend to have similar results. Standard errors were also of similar magnitude, with the exception of estimates when s_2 is LD2 where the standard errors for the SNM tended to be uninformative. All three methods agree on statistical significance for the effect of NO_3 when s_2 is the LD3 sampling location, 109 kilometers upstream of LD1. The point estimates at this location for NO_3 were 1.12 for the g-formula and 1.67 for the MSM and SNM, implying a 2- to 3-fold increase in LD1 chlorophyll when NO_3 is above 0.38 mg/L at both the location 109km upstream (LD3) and the location 132km upstream (near Cross Creek waste water treatment plant). Causal effects of NH_3, P and TKN consistently hover near zero with two exceptions. The effect of NH_3 appears to decrease after 109km upstream, and the effect of TKN appears to increase after 49km upstream.

Section 2.11 includes summaries of results for all cutpoints, space-varying confounding covariates, test statistic distribution settings, as well as multiple outcome and exposure model specifications. Point estimates varied modestly depending on the cutpoint, space-varying confounding covariate, and how the exposure/outcome models were specified. All of the point estimates for NO_3 were greater than zero for locations 109 and 95 kilometers upstream. Statistical significance was sensitive to the choice of Wald test statistic distribution but generally accords with the shifts in significance seen in the simulation results.

2.7 Discussion

Our results corroborate existing evidence that the much more abundant nitrate form of N is a major driver of downstream chlorophyll production in the middle section of the Cape Fear River. Among dissolved inorganic N species, cyanobacteria prefer to assimilate N in the form of ammonium and then switch to nitrate uptake when ammonium is depleted (Burkholder, 2002). In this section of the Cape Fear River, ammonium is typically at low concentrations while nitrate concentrations are an order of magnitude higher (Mallin et al., 2006; Kennedy and Whalen, 2007). Experimental additions of inorganic and organic N have stimulated algae growth in the Cape Fear River (Dubbs and Whalen, 2008) and its two major tributaries, the Black and Northeast Cape Fear Rivers (Mallin et al., 2004). As cyanobacteria are a primary harmful algal taxa group of concern in this system, it is notable that N stimulates growth of this group (Burkholder, 2002) as well as growth of *Microcystis* specifically (Paerl, 1987; Siegel et al., 2011; Yuan et al., 2014).

Both point and nonpoint sources such as agricultural runoff contribute to nutrient concentrations in the Cape Fear River (Rajbhandari et al., 2015). Our data cannot distinguish between sources of pollution. While the data also cannot precisely pinpoint locations of nutrient inputs into the river, our analysis does indicate areas for further investigation. Across our choices of cutpoints and various exposure and outcome models (see Section 2.11 for details), significant effects of NO_3 tended to occur between 86 and 109km upstream of LD1. Notably, four major National Pollutant Discharge Elimination System permits are located in this reach of the river: the Tarheel Plant (NC0078344), Dupont Fayetteville Works (NC0003573), Cedar Creek Site (NC0003719), and the Rockville Creek Waste Water Treatment plant (NC0050105).

We have shown how "what if" questions on water quality of scientific and policy interest can be mathematically framed as estimands. When the estimand involves space- and/or time-varying confounding, this must be accounted for in estimation, else biased estimates will result. Our application is one of the first to use the g-methods (Robins and Hernán, 2009) as part of an ecological causal assessment. In fact, despite their general utility, structural nested models have rarely been applied in practice (Vansteelandt and Joffe, 2014). We demonstrate how they can be implemented and give details for deriving a closed form estimator in Section 2.10.

Results from observational studies can always be skeptically reviewed, but the potential outcomes framework forms a basis for constructive critique. The causal assumptions described in Section 2.4.1 must be thoroughly vetted. Has all confounding been accounted for? Are the parametric forms of our models correctly specified, or reasonably so? As with any research, the causal effects estimated from a single analysis should not be the sole source of evidence. The approach in this chapter can be used to augment weight-of-evidence approaches such as the EPA's CADDIS. The potential outcomes framework is well suited to aid policymakers in development of permit standards and surface water standards. Yuan (2010), for example, estimated effects of nutrient pollution on stream invertebrates using propensity score methods.

Lastly, while GEE methods are available in several R packages (Halekoh et al., 2006; Carey, 2015), generic M-estimation is not straightforward with current statistical software. The causal models implemented in this chapter involve combining estimating equations from several models. The R package geex was developed in conjunction with this research to streamline a programmer's work in implementing estimating equation theory. The Supplementary Material includes the code to replicate our analyses, and examples of geex for causal models may be found therein.

2.8 Simulation details

The nodes in the simulated study system were parameterized and generated according to the following distributions:

$$t = 0 \begin{cases} L_{110} & \sim N(21.5, 2.5) \\ L_{220} & \sim N(-2.8, 0.7) \\ A_{s0} & \sim Bern(0.1)\text{for } s = 1, 2 \\ Y_{30} & \sim N(2.25, 1.25) \end{cases}$$

$$t = 1, 2, 3 \begin{cases} L_{11t} & \sim N(23 + 0.2L_{11,t-1}, 2) \\ A_{1t} & \sim Bern(\text{logit}^{-1}(-2.5 + 0.09L_{11t} + 0.025A_{1,t-1})) \\ L_{12t} & \sim N(6.75 + 0.75L_{11,t-1}, 1) \\ L_{22t} & \sim N(2 - 0.04L_{12t} + 0.04L_{22,t-1} + 0.3A_{1t}, 0.25) \\ A_{2t} & \sim Bern(\text{logit}^{-1}(-2.5 + 0.09L_{12t} + 0.1L_{22t} + 0.05A_{1t} + 0.025A_{2,t-1})) \\ Y_{3t} & \sim N(-5 + 1A_{s-1,t} + 0.5A_{s-2,t} + 0.025L_{1,s-1,t} + \\ & \quad 0.5L_{2,s-1,t} + 0.35Y_{3,t-1}, 1) \end{cases}$$

Code for the simulations can be found in the updown R package available upon request.

2.9 Parametric g-formula formulation

In all of the analyses, the outcome model was parameterized within the g-formula as a linear model:

$$E\left[Y_{3t}|\overline{A}_t = [0_{t-1}, a_t], \overline{O}_t = \overline{o}_t\right] = h(\overline{A}, \overline{O}, \beta) \tag{2.G1}$$

For example, the correctly specified h for the simulations is

$$\beta_0^g + \beta_1^g a_{2t} + \beta_2^g a_{1t} + \beta_3^g l_{12t} + \beta_4^g l_{22t} + \beta_5^g y_{3,t-1}.$$

The sensitivity analyses varied which L covariates were included in h, but the parameterization of the treatment terms (a_{2t} and a_{1t}) was not modified.

The average potential outcome $E\left[Y_{3t}([0_{t-1}, a_t])\right]$ can be linked to observed data in the following manner:

$$
\begin{aligned}
&E\left[Y_{3t}([0_{t-1}, a_t])\right] && \text{(2.G2)}\\
&= E\left\{E\left[Y_{3t}([0_{t-1}, a_t])|\overline{A}_{2t} = [0_{t-1}, a_t], \overline{L}_t = \overline{l}_t\right]\right\} \text{ (no unmeasured confounders)}\\
&= E\left\{E\left[Y_{3t}|\overline{A}_t = [0_{t-1}, a_t], \overline{L}_t = \overline{l}_t\right]\right\} \text{ (causal consistency)}
\end{aligned}
$$

According to the parametric g-formula, models for each $f_{l_{pjk}} = f_{l_{pjk}|\bar{l}_{j-1,k-1},\bar{a}_{j-1,k-1}}$ must be fit. However, as will be clear below, the parameters corresponding to non-space- or time-varying covariates cancel in a causal contrast. Hence, we only fit a model for $f_{l_{22t}}$, for which we used a standard linear model with expectation $\gamma_0^g + \gamma_1^g l_{12t} + \gamma_2^g l_{22,t-1} + \gamma_3^g a_{1t}$. By causal consistency, $E[L_{22t}(a_{1t})] = \gamma_0^g + \gamma_1^g l_{12t} + \gamma_2^g l_{22,t-1} + \gamma_3^g a_{1t}$. Under this assumed parameterization, $E[L_{22t}(a_{1t}) - L_{22t}(0)] = \gamma_3^g$.

Putting (2.G2) together with the model for $f_{l_{22t}}$ obtains:

$$
\begin{aligned}
&E\left[Y_{3t}([0_{t-1}, a_t]) - Y_{3t}(0_t)\right]\\
&= E\left\{E\left[Y_{3t}([0_{t-1}, a_t]) - Y_{3t}(0_t)\right]\middle| A_{1t} = 0, A_{2t} = [0_{t-1}, a_t], \overline{L}_{2t}\right\}\\
&= E\left\{\beta_1^g a_{2t} + \beta_2^g a_{1t} + \beta_4^g[L_{22t}(a_{1t}) - L_{22t}(0)]\right\} \text{ (plugging in } h)\\
&= \beta_1^g a_{2t} + \beta_2^g a_{1t} + \beta_4^g E\left\{E[L_{22t}(a_{1t}) - L_{22t}(0)|A_{11} = a_{11}, L_{12t}, L_{22,t-1}]\right\}\\
&= \beta_1^g a_{2t} + \beta_2^g a_{1t} + \beta_4^g \gamma_3^g.
\end{aligned}
$$

When $a_t = (1,1)'$ for all t, $\mu^g = \frac{1}{T}\sum_{t=1}^{T} E\left[Y_{3t}([0_{t-1}, a_t]) - Y_{3t}(0_t)\right] = \frac{1}{T}\sum_{t=1}^{T}[\beta_1^g + \beta_2^g + \beta_4^g\gamma_3^g]$.

2.10 Closed form estimator for SNM parameters

Vansteelandt and Joffe (2014) show that a consistent estimator of β^s can found by solving estimating equations (eq. 33):

$$\sum_i \sum_t \sum_s \{d_{st}(\overline{L}_{ist}, \overline{A}_{ist}) - E[d_{st}(\overline{L}_{ist}, \overline{A}_{ist})|\overline{L}_{ist}, \overline{A}_{i,s-1,t-1}]\}$$
$$\{U_{ist}(\beta^s) - E[U_{ist}(\beta^s)|\overline{L}_{ist}, \overline{A}_{i,s-1,t-1}]\}$$

where $d_{st}(\overline{L}_{ist}, \overline{A}_{ist})$ is chosen to be $E[\partial U_{st}(\beta^s)/\partial\beta^s|\overline{L}_{ist}, \overline{A}_{ist}]$. This formulation is slightly different from Vansteelandt and Joffe in that we added an additional dimension s. Since our endogenous covariate is space-varying rather than time-varying, the blip process is indexed by s rather than t.

Let $\rho_{ist} = E[d_{st}(\overline{L}_{ist}, \overline{A}_{ist})|\overline{L}_{ist}, \overline{A}_{i,s-1,t-1}]$ and $\lambda_{ist} = E[U_{ist}(\beta^s)|\overline{L}_{ist}, \overline{A}_{i,s-1,t-1}]$, then:

$$
\begin{aligned}
&\sum_i \sum_t \big\{[d_{1t}(\overline{L}_{i1t}, \overline{A}_{i1t}) - \rho_{i1t}](U_{i1t}(\beta^s) - \lambda_{i1t}) + \\
&\qquad\qquad [d_{2t}(\overline{L}_{i2t}, \overline{A}_{i2t}) - \rho_{i2t}](U_{i2t}(\beta^s) - \lambda_{i2t})\big\} \\
&= \sum_i \sum_t \left\{ \left[\begin{pmatrix} A_{i2t} \\ 0 \end{pmatrix} - \rho_{i2t} \right] (U_{i1t}(\beta^s) - \lambda_{i1t} + \left[\begin{pmatrix} A_{2t} \\ A_{2t} \end{pmatrix} - \rho_{i2t} \right] (U_{i2t}(\beta^s) - \lambda_{i2t}) \right\} \\
&= \sum_i \sum_t \left\{ \left[\begin{pmatrix} A_{i2t} \\ 0 \end{pmatrix} - \rho_{i2t} \right] (Y_{i3t} - \beta^s_1 A_{i2t} - \lambda_{i1t}) \right. \\
&\qquad\quad \left. + \left[\begin{pmatrix} A_{i2t} \\ A_{i2t} \end{pmatrix} - \rho_{i2t} \right] (Y_{i3t} - \beta^s_1 A_{i2t} - \beta^s_2 A_{i1t} - \lambda_{i2t}) \right\} \\
&= \sum_i \sum_t \left\{ \begin{pmatrix} B^0_{2t}(r_{i3t} - \beta^s_1 A_{2t}) + B^1_{2t}(r^0_i - \beta^s_1 A_{2t} - \beta^s_2 A_{1t}) \\ B^1_{1t}(r^1_i - \beta^s_1 A_{2t} - \beta^s_2 A_{1t}) \end{pmatrix} \right\}.
\end{aligned}
$$

In the last line, we let $B_{ist} = A_{2t} - \rho^k_{ist}$. Let $r_{ist} = Y_{i3t} - \lambda^k_{ist}$.

Let $C_i = \sum_t r^0_i (B^0_{i2t} + B^1_{i2t})$, $D_i = \sum_t (B^0_{i2t} A_{i2t} + B^1_{i2t} A_{i2t})$, $E_i = \sum_t B^1_{i2t} A_{i1t}$, $F_i = \sum_t B^1_{1t} r^1_i$, $G_i = B^1_{i1t} A_{i2t}$, and $H_i = \sum_t B^1_{i1t} A_{i1t}$. Then β^s is the solution to:

$$
\sum_i \begin{pmatrix} C_i - \beta^s_1 D_i - \beta^s_2 E_i \\ F_i - \beta^s_1 G_i - \beta^s_2 H_i \end{pmatrix}
$$

which yields,

$$
\hat{\beta}^s_1 = \frac{\sum_i E_i \sum_i F_i - \sum_i C_i \sum_i H_i}{\sum_i E_i \sum_i G_i - \sum_i D_i \sum_i H_i} \text{ and } \hat{\beta}^s_2 = \frac{\sum_i D_i \sum_i F_i - \sum_i C_i \sum_i G_i}{\sum_i D_i \sum_i H_i - \sum_i E_i \sum_i G_i}
$$

where

$$\sum_i E_i \sum_i G_i \neq \sum_i D_i \sum_i H_i, \quad \sum_i E_i \neq 0.$$

2.11 Stability analyses

In addition to estimating the target parameters using all possible combinations of settings of the cutpoint and space-varying confounding variables, we also modified the exposure and treatment models to include a temperature by flow interaction term in both outcome and exposure models. This resulted in a total of 1200 point estimates per method. If some component model failed to converge for a method, then the estimate attempt was considered a failure. Across all 3600 attempts, model fitting failed 272 times for the MSMs, 70 times for the SNMs, and zero times for the g-formula.

Figure 2.9 shows all the point estimates within (-6, 6) across all the model options. In two cases, estimates from an SNM were outside this range. The results in Figure 2.9 conform to the general patterns described in the main text.

Figure 2.9: This figure shows the 3258 point estimates obtained across settings of the cutpoint, space-varying confounding variables, and exposure and outcome models. Point estimates for NO_3 between 86 and 109 km upstream are positive across all settings and have the most consistently strong effects. TKN estimates for the two locations just upstream of LD1 are all positive, while the P estimates from the same locations are all negative.

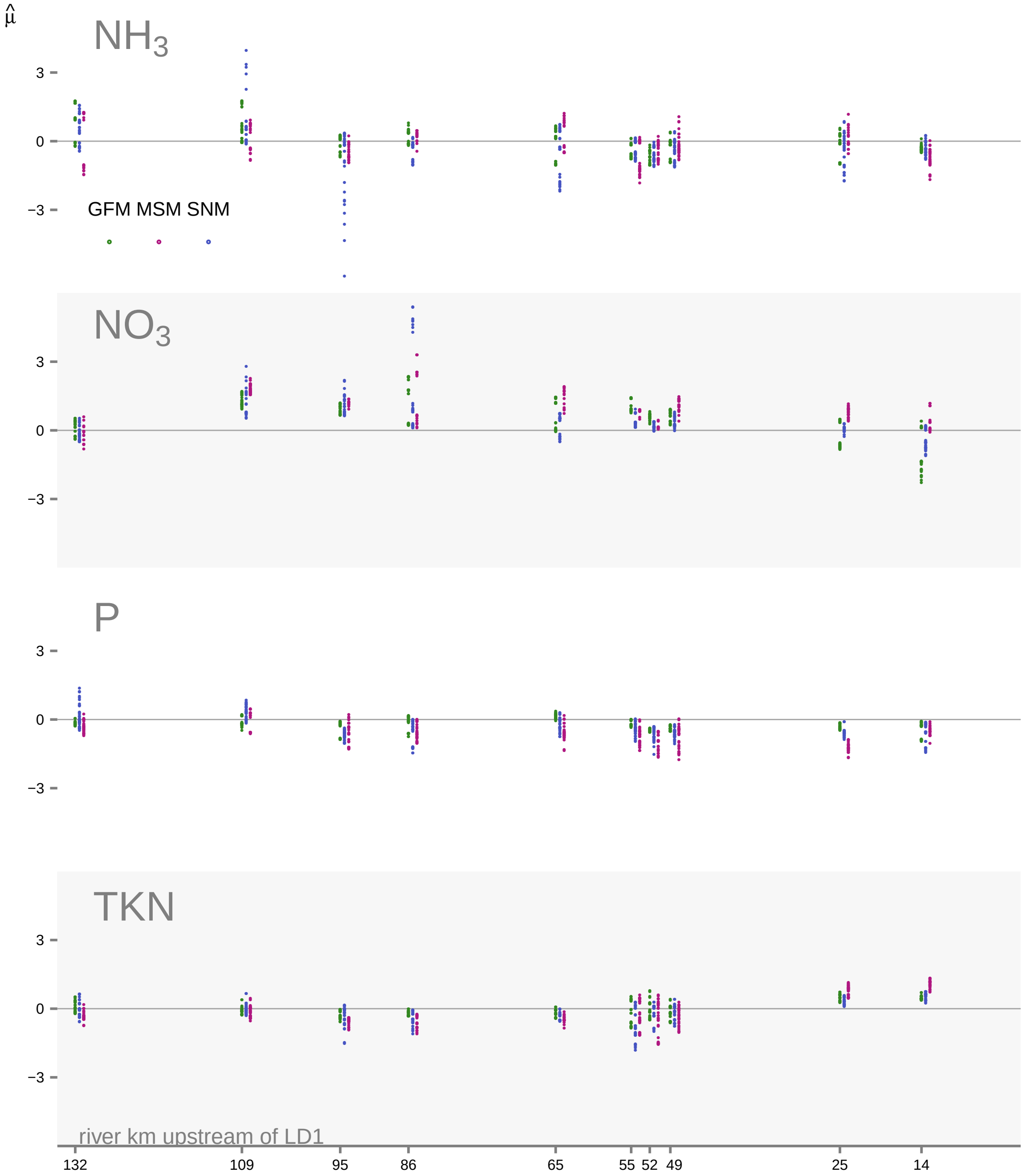
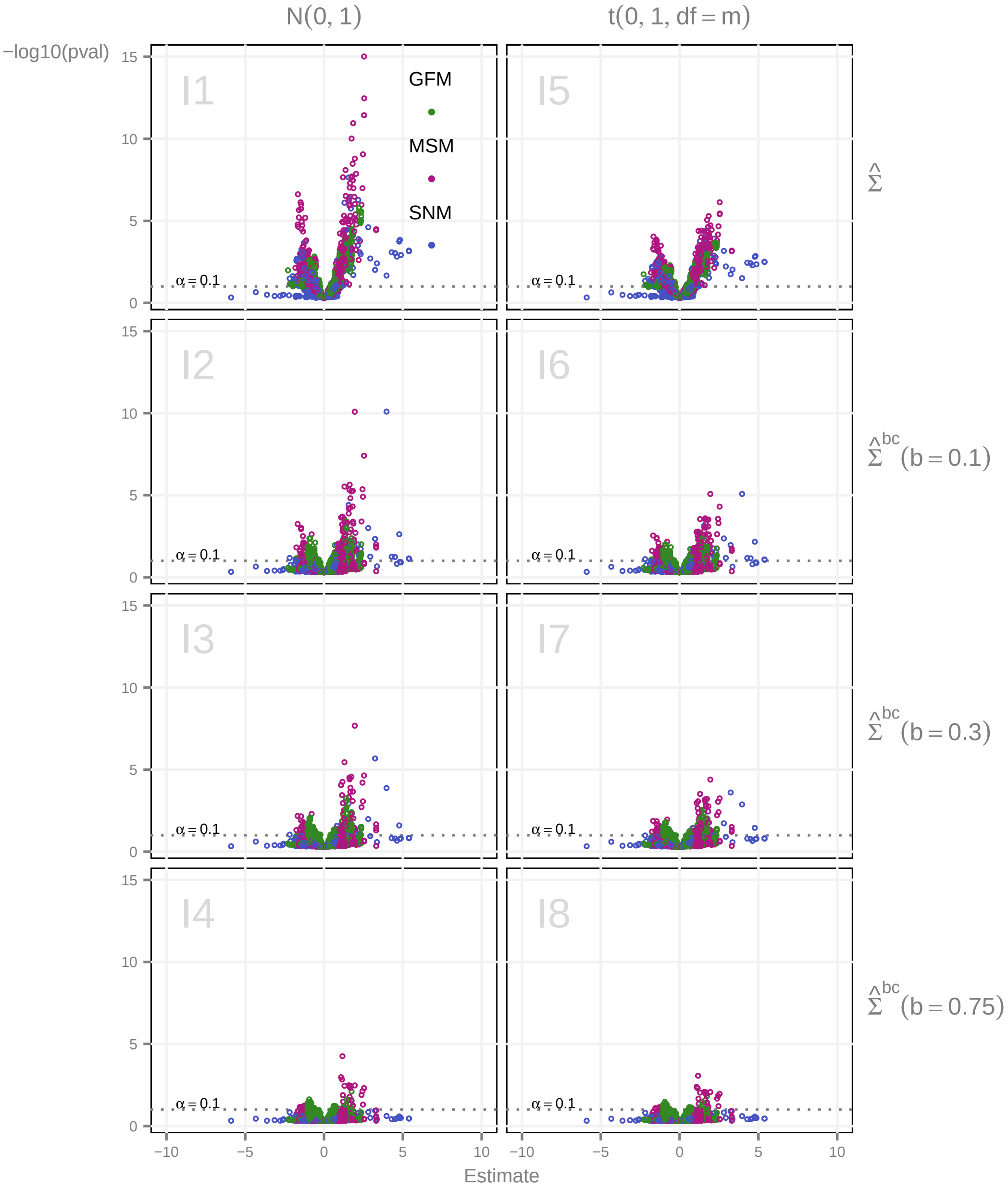


Figure 2.10 show all the point estimates along with the p-values using different test statistic distributions. In each panel, the point estimates are the same, but the significance clearly depends on the test statistic distribution.

Figure 2.10: Volcano plot of estimates of the Cape Fear analysis showing significance and estimates for all analysis methods and models. The point estimates are same in all the panels. Significance levels vary depending on the distribution used for the test statistic.



CHAPTER 3: THE CALCULUS OF M-ESTIMATION USING GEEX

3.1 Introduction

M-estimation, or estimating equation, methods are a simple but general class of statistical procedures for carrying out point estimation and asymptotic inference (Boos and Stefanski, 2013). Originally developed in the context of studying the large sample properties of robust statistics (Huber and Ronchetti, 2009), M-estimation is applicable in a broad range of settings. The general result from M-estimation theory states that if an estimator can be expressed as the solution to an unbiased estimating equation, then under suitable regularity conditions the estimator is asymptotically Normal and the asymptotic variance of the estimator can be consistently estimated using the empirical sandwich estimator. Many estimators can be expressed as solutions to unbiased estimating equations, thus M-estimation has extensive applicability. The primer by Stefanski and Boos (2002) demonstrates how a wide array of statistics can be expressed as M-estimators, including the popular method of generalized estimating equations (GEE) (Liang and Zeger, 1986) for longitudinal data analysis.

Despite the broad applicability of M-estimation, existing statistical software packages implement M-estimators for particular forms of estimating equations. This paper introduces the R (R Core Team, 2016) package `geex`, a general M-estimation programming framework that can obtain point and variance estimates from any set of unbiased estimating equations. The analyst defines a function that takes unit-level data and returns a function in terms of parameters. From this single function, `geex` then uses numerical routines to compute parameter point and variance estimates.

Below, Section 3.2 reviews M-estimation theory and outlines how `geex` translates mathematical expressions of estimating functions into code. Section 3.3 implements three examples from Stefanski and Boos (2002) (hereafter SB) using `geex` plus an example of GEE. All of the SB examples and several more are available at the package website (`https://bsaul.github.io/geex/`). Section 3.4 compares `geex` to existing R packages. Section 3.5 demonstrates the finite sample correction feature of `geex`, and Section 3.6 concludes with a brief discussion of the software's didactic utility and pragmatic applications.

3.2 From M-estimation math to code

In the basic set-up, M-estimation applies to estimators of the $p \times 1$ parameter θ which can be obtained as solutions to an equation of the form

$$\sum_{i=1}^{m} \psi(O_i, \theta) = 0, \tag{3.7}$$

where $O_1, \dots, O_m$ are m independent and identically distributed (iid) random variables, and the function ψ returns a vector of length p and does not depend on i or m. See Stefanski and Boos (2002) for the case where the O_i are independent but not necessarily identically distributed. The roots of (1) are referred to as M-estimators and denoted by $\hat{\theta}$. M-estimators can be solved for analytically in some cases or computed numerically in general. Under certain regularity conditions, the asymptotic properties of $\hat{\theta}$ can be derived from Taylor series approximations, the law of large numbers, and the central limit theorem (Boos and Stefanski, 2013, sec. 7.2). In brief, let θ_0 be the true parameter value defined by $\int \psi(o, \theta_0) dF(o) = 0$, where F is the distribution function of O. Let $\dot{\psi}(o, \theta) = \partial\psi(O_i, \theta)/\partial\theta^{\intercal}$, $A(\theta_0) = E[-\dot{\psi}(O_1, \theta_0)]$, and $B(\theta_0) = E[\psi(O_1, \theta_0)\psi(O_1, \theta)^{\intercal}]$. Then under suitable regularity assumptions, $\hat{\theta}$ is consistent and asymptotically Normal, i.e.,

$$\sqrt{m}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0, V(\theta_0)) \text{ as } m \to \infty,$$

where $V(\theta_0) = A(\theta_0)^{-1}B(\theta_0)\{A(\theta_0)^{-1}\}^\intercal$. The sandwich form of $V(\theta_0)$ suggests several possible large sample variance estimators. For some problems, the analytic form of $V(\theta_0)$ can be derived and estimators of θ_0 and other unknowns simply plugged into $V(\theta_0)$. Alternatively, $V(\theta_0)$ can be consistently estimated by the empirical sandiwch variance estimator, where the expectations in $A(\theta)$ and $B(\theta)$ are replaced with their empirical counterparts. Let $A_i = -\dot{\psi}(O_i, \theta)|_{\theta=\hat{\theta}}$, $A_m = \sum_{i=1}^m A_i/m$, $B_i = \psi(O_i, \hat{\theta})\psi(O_i, \hat{\theta})^\intercal$, and $B_m = \sum_{i=1}^m B_i/m$. The empirical sandwich estimator of the variance of $\hat{\theta}$ is:

$$\hat{\Sigma} = A_m^{-1}B_m\{A_m^{-1}\}^\intercal/m. \tag{3.8}$$

The `geex` package provides an application programming interface (API) for carrying out M-estimation. The analyst provides a function, denoted `estFUN`, corresponding to $\psi(O_i, \theta)$ that maps data O_i to a function of θ. Numerical derivatives approximate $\dot{\psi}$ so that evaluating $\hat{\Sigma}$ is entirely a computational exercise. No analytic derivations are required from the user.

Consider estimating the population mean $\theta = E[Y_i]$ using the sample mean $\hat{\theta} = m^{-1}\sum_{i=1}^m Y_i$ of m iid random variables $Y_1, \dots, Y_m$. The estimator $\hat{\theta}$ can be expressed as the solution to the following estimating equation:

$$\sum_{i=1}^m (Y_i - \theta) = 0.$$

which is equivalent to solving (3.7) where that $O_i = Y_i$ and $\psi(O_i, \theta) = Y_i - \theta$. An `estFUN` is a translation of ψ into a function whose first argument is `data` and returns a function whose first argument is `theta`. An `estFUN` corresponding to the estimating equation for the sample mean of Y is:

```
meanFUN <- function(data){ function(theta){ data$Y - theta } } .
```

The `geex` package exploits R as functional programming language: functions can return and modify other functions (Wickham, 2014, ch. 10). If an estimator fits into the above framework, then the user need only correctly specify `estFUN`. No other programming is required to obtain point and variance estimates. The remaining sections provide examples of translating ψ into an `estFUN`.

3.3 Calculus of M-estimation Examples

The package can be installed from CRAN with `install.packages("geex")`. The first three examples of M-estimation from SB are presented here for demonstration. For these examples, the data are $O_i = \{Y_{1i}, Y_{2i}\}$ where $Y_1 \sim N(5, 16)$ and $Y_2 \sim N(2, 1)$ for $m = 100$ and are included in the `geexex` dataset. The last example in this section demonstrates a GEE application, which is elaborated on in Section 3.5 to demonstrate finite sample corrections.

3.3.1 Example 1: Sample moments

The first example estimates the population mean (θ_1) and variance (θ_2) of Y_1. Figure 3.11 shows the estimating equations and corresponding `estFUN` code. The solution to the estimating equations in Figure 3.11 are $\hat{\theta}_1 = m^{-1}\sum_{i=1}^m Y_{1i}$ and $\hat{\theta}_2 = m^{-1}\sum_{i=1}^m (Y_{1i} - \hat{\theta}_1)^2$, i.e., the sample mean and variance.

Figure 3.11: Estimating equations and `estFUN` for example 1

$$\psi(Y_{1i}, \theta) = \begin{pmatrix} Y_{1i} - \theta_1 \\ (Y_{1i} - \theta_1)^2 - \theta_2 \end{pmatrix}$$

```
SB1_estfun <- function(data){
  Y1 <- data$Y1
  function(theta){
      c(Y1 - theta[1],
        (Y1 - theta[1])^2 - theta[2])
  }
}
```

The function `m_estimate` can compute both $\hat{\theta}$ and $\hat{\Sigma}$. Two inputs are required for `m_estimate`: `estFUN` (the ψ function), `data` (the analysis data frame containing O_i for $i = 1, \ldots, m$). The package defaults to `rootSolve::multiroot` (Soetaert and Herman, 2009; Soetaert, 2009) for estimating roots of (3.7), though the solver algorithm can be specified in the `root_control` argument. Starting values for `multiroot` are passed via the `root_control` argument.

```
library(geex)
results <- m_estimate(
    estFUN = SB1_estfun,
    data  = geexex,
    root_control = setup_root_control(start = c(1,1)))
```

Table 3.1: Results obtained from geex and closed form estimators

	geex		Closed form	
$\hat{\theta}$	5.045	10.041	5.045	10.041
$\hat{\Sigma}$	0.100	0.037	0.100	0.037
	0.037	2.492	0.037	2.492

The `m_estimate` function returns an object of the `S4` class `geex`, which contains an `estimates` slot and `vcov` slot for $\hat{\theta}$ and $\hat{\Sigma}$, respectively. These slots can be accessed by the functions `coef` (or `roots`) and `vcov`. The point estimates obtained for θ_1 and θ_2 are analogous to the base R functions `mean` and `var` (after multiplying by $m - 1/m$). SB gave a closed form for $A(\theta_0)$ (an identity matrix) and $B(\theta_0)$ (not shown) and suggest plugging in sample moments to compute $B(\hat{\theta})$. The point and variance estimates using both `geex` and the analytic solutions are shown in Table 3.1. The maximum absolute difference between either the point or variance estimates is 4.3e-11, thus demonstrating excellent agreement between the numerical results obtained from `geex` and the closed form solutions for this set of estimating equations and data.

3.3.2 Example 2: Ratio estimator

This example calculates a ratio estimator (Figure 3.12) and illustrates the delta method via M-estimation. The estimating equations target the means of Y_1 and Y_2, labelled θ_1 and θ_2, as well as the estimand $\theta_3 = \theta_1/\theta_2$.

Figure 3.12: Estimating equations and `estFUN` for example 2

$$\psi(O_i, \theta) = \begin{pmatrix} Y_{1i} - \theta_1 \\ Y_{2i} - \theta_2 \\ \theta_1 - \theta_3\theta_2 \end{pmatrix}$$

```
SB2_estfun <- function(data){
  Y1 <- data$Y1; Y2 <- data$Y2
  function(theta){
      c(Y1 - theta[1],
        Y2 - theta[2],
        theta[1] - (theta[3]*theta[2])
    )
  }
}
```

The solution to (3.7) for this ψ function yields $\hat{\theta}_3 = \bar{Y}_1/\bar{Y}_2$, where $\bar{Y}_j$ denotes the sample mean of $Y_{j1}, \ldots, Y_{jm}$ for $j = 1, 2$.

SB gave closed form expressions for $A(\theta_0)$ and $B(\theta_0)$, into which we plug in appropriate estimates for the matrix components and compare to the results from `geex`. The point estimates again show excellent agreement (maximum absolute difference = 4.4e-16), while the covariance estimates differ by the fourth decimal (maximum absolute difference = 1e-03).

3.3.3 Example 3: Delta method continued

This example extends Example 1 to again illustrate the delta method. The estimating equations target not only the mean (θ_1) and variance (θ_2) of Y_1, but also the standard deviation (θ_3) and the log of the variance (θ_4) of Y_1.

Figure 3.13: Estimating equations and `estFUN` for example 3

$$\psi(Y_{1i}, \theta) = \begin{pmatrix} Y_{1i} - \theta_1 \\ (Y_{1i} - \theta_1)^2 - \theta_2 \\ \sqrt{\theta_2} - \theta_3 \\ \log(\theta_2) - \theta_4 \end{pmatrix}$$

```
SB3_estfun <- function(data){
  Y1 <- data$Y1
  function(theta){
      c(Y1 - theta[1],
       (Y1 - theta[1])^2 - theta[2],
       sqrt(theta[2]) - theta[3],
       log(theta[2]) - theta[4])
  }
}
```

SB again provided a closed form for $V(\theta_0)$, which we compare to the `geex` results. Three of the point estimates are identical up to machine epsilon, while the $\hat{\theta}_4$ estimates are equivalent up to the 15th decimal. The covariance estimates have a maximum absolute difference of 3.8e-11.

3.3.4 Example 4: Generalized Estimating Equations

In their seminal paper, Liang and Zeger (1986) introduced the following class of estimating equations for the analysis of longitudinal or clustered data:

$$\sum_{i=1}^{m} \psi(\mathbf{O}_i, \beta) = \sum_{i=1}^{m} \mathbf{D}_i^{\intercal} \mathbf{V}_i^{-1} [\mathbf{Y}_i - \mu(\mathbf{X}_i; \beta)] = 0 \tag{3.9}$$

where $\mathbf{O}_i = \{\mathbf{Y}_i, \mathbf{X}_i\}$ and $\mathbf{D}_i = \partial\mu/\partial\beta$. The sample size m is the number of independent clusters and each cluster has size n_i, so that $\mathbf{Y}_i$ is a $n_i \times 1$ vector and $\mathbf{X}_i$ is a $n_i \times p$ matrix. The covariance matrix is modeled by $\mathbf{V}_i = \phi \mathbf{A}_i^{0.5} \mathbf{R}(\alpha) \mathbf{A}_i^{0.5}$. The matrix $\mathbf{R}(\alpha)$ is the "working" correlation matrix. The example uses an exchangeable correlation structure with off-diagonal elements α. The matrix $\mathbf{A}_i$ is a matrix with diagonal elements containing the variance functions of μ. The equations in (3.9) can be translated into an `estFUN` as:

```
gee_estfun <- function(data, formula, family){
  X <- model.matrix(object = formula, data = data)
  Y <- model.response(model.frame(formula = formula, data = data))
  n <- nrow(X)
  function(theta, alpha, psi){
    mu  <- family$linkinv(X %*% theta)
    Dt  <- t(X) %*% diag(as.numeric(mu), nrow = n)
    A   <- diag(as.numeric(family$variance(mu)), nrow = n)
    R   <- matrix(alpha, nrow = n, ncol = n)
    diag(R) <- 1
    V   <- psi * (sqrt(A) %*% R %*% sqrt(A))
    Dt %*% solve(V) %*% (Y - mu)
  }
}
```

This estFUN treats the correlation parameter α and scale parameter ϕ as fixed, though some estimation algorithms use an iterative procedure that alternates between estimating β and these parameters. By customizing the root finding function, such an algorithm can be implemented using geex [see vignette("v03_root_solvers") for more information].

We use this example to compare covariance estimates obtained from the gee function (Carey, 2015), so root finding computations of geex are turned off. To compute only the sandwich variance estimator, set compute_roots = FALSE and pass estimates of θ_0 via the roots argument (see the example in Section 3.5). The gee β estimates are used for this example. Estimates for α and ϕ are also extracted from the gee results in m_estimate. This example shows that an estFUN can accept additional arguments to be passed to either the outer (data) function or the inner (theta) function. Unlike previous examples, the independent units are clusters (types of wool), which is specified in m_estimate by the units argument. This argument defaults to NULL, in which case m equals the number of rows in the data frame.

```
g <- gee::gee(breaks~tension, id=wool, data=warpbreaks,
              corstr="exchangeable")
```

```
results <- m_estimate(
  estFUN = gee_estfun,
  data  = warpbreaks,
  units = "wool",
  roots = coef(g),
  compute_roots = FALSE,
  outer_args = list(formula = breaks ~ tension,
                    family  = gaussian()),
  inner_args = list(alpha   = g$working.correlation[1,2],
                    psi     = g$scale))
```

The maximum absolute difference between the estimated covariances computed by gee and geex is 1.3e-09.

3.4 Comparison to existing software

The above examples demonstrate the basic utility of the geex package and the power of R's functional programming capability. To our knowledge, geex is the first R package to create an extensible API for any estimator that satisfies (3.7). Existing R packages such as gee (Carey, 2015), geepack (Halekoh et al., 2006), and geeM (McDaniel et al., 2013) solve for parameters in a GEE framework. Other packages such as fastM (Duembgen et al., 2014) and smoothmest (Hennig, 2012) implement M-estimators for specific use cases.

For computing a sandwich variance estimator, geex is similar to the popular sandwich package (Zeileis, 2004, 2006), which computes the empirical sandwich variance estimator from modelling methods such as lm, glm, gam, survreg, and others. For comparison to the exposition herein, the infrastructure of sandwich is explained in Zeileis (2006). Advantages of geex compared to sandwich include: (i) for custom applications, a user only needs to specify a single estFUN function as opposed to both the bread and estfun functions; (ii) as demonstrated in the examples above, the syntax of an estFUN may closely resemble the mathematical expression of the corresponding estimating function; (iii) estimating functions from multiple models are easily stacked; and (iv) point estimates can be obtained. The precision and computational speed of point and variance estimation in geex, however, depends on numerical approximations rather than analytic expressions.

An example compares sandwich and geex in the goal of estimating $\hat{\Sigma}$ from a simple linear model. Consider estimating $\hat{\Sigma}$ for the β parameters in the following model contained in the geexex data: $Y_4 = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$, where $\epsilon \sim N(0, 1)$. The estimating equation for this model can be expressed in an estFUN as:

```
lm_estfun <- function(data){
  X <- cbind(1, data$X1, data$X2)
  Y <- data$Y4
  function(theta){
    t(X) %*% (Y - X %*% theta)
  }
}
```

Then $\hat{\beta}$ and $\hat{\Sigma}$ can be computed in geex:

```
results <- m_estimate(
  estFUN = lm_estfun,
  data  = geexex,
  root_control = setup_root_control(start = c(0, 0, 0)))
```

or from the lm and sandwich functions:

```
fm <- lm(Y4 ~ X1 + X2, data = geexex)
sand_vcov <- sandwich::sandwich(fm)
```

The results are virtually identical numerically (maximum absolute difference: 1.4e-12. The lm/sandwich option is faster computationally, but this difference can be reduced by, for example, changing the options of the derivative function via deriv_control or computing $\hat{\Sigma}$ using the parameter estimates from lm. The important difference is that geex lays bare the estimating function used, which is both a powerful didactic tool as well as a programming advantage when developing custom estimating functions.

3.5 Finite sample corrections

The empirical sandwich variance estimator is known to underestimate $V(\theta_0)$ in small samples (Fay and Graubard, 2001). Particularly in the context of GEE, many authors (e.g., see Kauermann and Carroll, 2001; Mancl and DeRouen, 2001; Lu et al., 2007; Teerenstra et al., 2010; Paul and Zhang, 2014; Li and Redden, 2015) have proposed corrections that modify components of $\hat{\Sigma}$ and/or by assuming $\hat{\theta}$ follows a t (or F), as opposed to Normal, distribution with some estimated degrees of freedom. Many of the proposed corrections somehow modify a combination of the A_i, A_m, B_i, or B_m matrices.

Users may specify functions that utilize these matrices to form corrections within `geex`. A finite sample correction function at a minimum takes the argument `components`, which is an `S4` object with slots for `A` ($= \sum_i A_i$) matrix, `A_i` (a list of all m A_i matrices), `B` ($= \sum_i B_i$) matrix, and `B_i` (a list of all m B_i matrices). Additional arguments may also be specified, as shown in the example. The `geex` package includes the bias correction and degrees of freedom corrections proposed by Fay and Graubard (2001) in the `fay_bias_correction` and `fay_df_correction` functions respectively. The following demonstrates the construction and use of the bias correction. Fay and Graubard (2001) proposed the modified variance estimator $\hat{\Sigma}^{bc}(b) = A_m^{-1} B_m^{bc}(b) \{A_m^{-1}\}^{\intercal}/m$, where:

$$B_m^{bc}(b) = \sum_{i=1}^{m} H_i(b) B_i H_i(b)^{\intercal},$$

$$H_i(b) = \{1 - \min(b, \{A_i A_m\}_{jj})\}^{-1/2},$$

and W_{jj} denotes the jj element of a matrix W. When $\{A_i A\}_{jj}$ is close to 1, the adjustment to $\hat{\Sigma}^{bc}(b)$ may be extreme, and the constant b is chosen by the analyst to limit over adjustments. The bias corrected estimator $\hat{\Sigma}^{bc}(b)$ can be implemented in `geex` by the following function:

```
bias_correction <- function(components, b){
  A   <- grab_bread(components)
  A_i <- grab_bread_list(components)
  B_i <- grab_meat_list(components)
  Ainv <- solve(A)

  H_i <- lapply(A_i, function(m){
    diag( (1 - pmin(b, diag(m %*% Ainv) ) )^(-0.5) )
  })

  Bbc_i <- lapply(seq_along(B_i), function(i){
    H_i[[i]] %*% B_i[[i]] %*% H_i[[i]]
  })
  Bbc   <- apply(simplify2array(Bbc_i), 1:2, sum)

  compute_sigma(A = A, B = Bbc)
}
```

The compute_sigma function simply computes $A^{-1}B\{A^{-1}\}^{\intercal}$. Note that geex computes A_m and B_m as the sums of A_i and B_i rather than the means, hence the appropriate function in the apply call is sum and not mean. To use this bias correction, the m_estimate function accepts a named list of corrections to perform. Each element of the list is a correct_control S4 object that can be created with the helper function correction, which accepts the argument FUN (the correction function) plus any arguments passed to FUN besides components.

```
results <- m_estimate(
  estFUN = gee_estfun, data  = warpbreaks,
  units = 'wool', roots = coef(g), compute_roots = FALSE,
  outer_args = list(formula = breaks ~ tension,
                    family  = gaussian()),
  inner_args = list(alpha   = g$working.correlation[1,2],
                    psi     = g$scale),
  corrections = list(
   bias_correction_.1 = correction(FUN =bias_correction, b = .1),
   bias_correction_.3 = correction(FUN =bias_correction, b = .3)))
```

In the geex output, the slot corrections contains a list of the results of computing each item in the corrections, which can be accessed with the get_corrections function. The corrections of Fay and Graubard (2001) are also implemented in the saws package (Fay and Graubard, 2001). Comparing the geex results to the results of the saws::geeUOmega function, the maximum absolute difference for any of the corrected estimated covariance matrices is 1.8e-09.

3.6 Summary

This chapter demonstrates how M-estimators and finite sample corrections can be transparently implemented in `geex`. Additional examples on the package website (`https://bsaul.github.io/geex/`) showcase the broad applicability of M-estimation including instrumental variables, sample quantiles, robust regression, generalized linear models, and more. A valuable feature of M-estimators is that estimating functions corresponding to multiple models may be combined, or "stacked," in a single set of estimating functions. The `geex` package makes it easy to stack estimating functions for the target parameters with estimating functions from each of the component models, as shown in the package vignette `v06_causal_example`. Indeed, the software was motivated by causal inference problems (Saul et al., 2017) where target causal parameters are functions of parameters in multiple models.

The theory of M-estimation is a broadly applicable statistical framework, yet existing R packages only implement particular classes of M-estimators. With its functional programming capabilities, R routines can be more general. The `geex` framework epitomizes the extensible nature of M-estimators and explicitly translates the estimating function ψ into a corresponding `estFUN`. In this way, `geex` should be useful for practitioners developing M-estimators, as well as students learning estimating equation theory.

3.7 Vignettes

As mentioned above, the package website contains several vignettes (tutorials) and additional articles about using `geex`. The vignette which demonstrates examples 4-10 of SB is included here as an example.

3.7.1 Example 4: Instrumental variables

Example 4 calculates an instumental variable estimator. The variables (Y_3, W_1, Z_1) are observed according to:

$$Y_{3i} = \alpha + \beta X_{1i} + \sigma_\epsilon \epsilon_{1,i}$$

$$W_{1i} = \beta X_{1i} + \sigma_U \epsilon_{2,i}$$

and

$$Z_{1i} = \gamma + \delta X_{1i} + \sigma_\tau \epsilon_{3,i}.$$

Y_3 is a linear function of a latent variable X_1, whose coefficient β is of interest. W_1 is a measurement of the unobserved X_1, and Z_1 is an instrumental variable for X_1. The instrumental variable estimator is $\hat{\beta}_{IV}$ is the ratio of least squares regression slopes of Y_3 on Z_1 and Y_3 and W_1. The ψ function below includes this estimator as $\hat{\theta}_4 = \hat{\beta}_{IV}$. In the example data, 100 observations are generated where $X_1 \sim \Gamma(\text{shape} = 5, \text{rate} = 1)$, $\alpha = 2$, $\beta = 3$, $\gamma = 2$, $\delta = 1.5$, $\sigma_\epsilon = \sigma_\tau = 1$, $\sigma_U = .25$, and $\epsilon_{\cdot,i} \sim N(0, 1)$.

$$\psi(Y_{3i}, Z_{1i}, W_{1i}, \theta) = \begin{pmatrix} \theta_1 - Z_{1i} \\ \theta_2 - W_{1i} \\ (Y_{3i} - \theta_3 W_{1i})(\theta_2 - W_{1i}) \\ (Y_{3i} - \theta_4 W_{1i})(\theta_1 - Z_{1i}) \end{pmatrix}$$

```
SB4_estFUN <- function(data){
  Z1 <- data$Z1; W1 <- data$W1; Y3 <- data$Y3
  function(theta){
      c(theta[1] - Z1,
        theta[2] - W1,
        (Y3 - (theta[3] * W1)) * (theta[2] - W1),
        (Y3 - (theta[4] * W1)) * (theta[1] - Z1)
    )
  }
}
```

```
estimates <- m_estimate(
  estFUN = SB4_estFUN,
  data  = geexex,
  root_control = setup_root_control(start = c(1, 1, 1, 1)))
```

The R packages `AER` and `ivpack` can compute the IV estimator and sandwich variance estimators respectively, which is done below for comparison.

```
ivfit <- AER::ivreg(Y3 ~ W1 | Z1, data = geexex)
iv_se <- ivpack::cluster.robust.se(ivfit,
                                   clusterid = 1:nrow(geexex))
```

```
coef(ivfit)[2]
```

```
##       W1
## 3.041264
```

```
coef(estimates)[4]
```

```
## [1] 3.041264
```

```
iv_se[2, 'Std. Error']
```

```
## [1] 0.01049669
```

```
sqrt(vcov(estimates)[4, 4])
```

```
## [1] 0.01039119
```

3.7.2 Example 5: Hodges-Lehmann estimator

This example shows the link between the influence curves and the Hodges-Lehman estimator.

$$\psi(Y_{2i}, \theta) = \left(IC_{\hat{\theta}_{HL}}(Y_2; \theta_0) - (\theta_1 - \theta_0) \right)$$

The functions used in `SB5_estFUN` are defined below:

```
F0 <- function(y, theta0, distrFUN = pnorm){
  distrFUN(y - theta0, mean = 0)
}

f0 <- function(y, densFUN){
  densFUN(y, mean = 0)
}

integrand <- function(y, densFUN = dnorm){
  f0(y, densFUN = densFUN)^2
}

IC_denom <- integrate(integrand, lower = -Inf, upper = Inf)$value


SB5_estFUN <- function(data){
  Yi <- data[['Y2']]
  function(theta){

      (1/IC_denom) * (F0(Yi, theta[1]) - 0.5)
  }
}


estimates <- m_estimate(
  estFUN = SB5_estFUN,
  data  = geexex,
  root_control = setup_root_control(start = 2))
```

The `hc.loc` of the `ICSNP` package computes the Hodges-Lehman estimator and SB gave closed form estimators for the variance.

```
theta_cls <- ICSNP::hl.loc(geexex$Y2)
Sigma_cls <- 1/(12 * IC_denom^2) / nrow(geexex)
```

```
## $geex
## $geex$parameters
## [1] 2.026376
##
## $geex$vcov
##             [,1]
## [1,] 0.01040586
##
##
## $cls
## $cls$parameters
## [1] 2.024129
##
## $cls$vcov
## [1] 0.01047198
```

3.7.3 Example 6: Huber center of symmetry estimator

This example illustrates estimation with nonsmooth ψ function for computing the Huber (1964) estimator of the center of symmetric distributions.

$$\psi_k(Y_{1i}, \theta) = \begin{cases} (Y_{1i} - \theta) & \text{when } |(Y_{1i} - \theta)| \leq k \\ k * sgn(Y_{1i} - \theta) & \text{when } |(Y_{1i} - \theta)| > k \end{cases}$$

```
SB6_estFUN <- function(data, k = 1.5){
  Y1 <- data$Y1
  function(theta){
    x <- Y1 - theta[1]
    if(abs(x) <= k) x else sign(x) * k
  }
}
```

```
estimates <- m_estimate(
  estFUN = SB6_estFUN,
  data  = geexex,
  root_control = setup_root_control(start = 3))
```

The point estimate from `geex` is compared to the estimate obtained from the `huber` function in the `MASS` package. The variance estimate is compared to an estimator suggested by SB: $A_m = \sum_i [-\psi'(Y_i - \hat{\theta})]/m$ and $B_m = \sum_i \psi_k^2(Y_i - \hat{\theta})/m$, where ψ_k' is the derivative of ψ where is exists.

```
theta_cls <- MASS::huber(geexex$Y1, k = 1.5, tol = 1e-10)$mu

psi_k <- function(x, k = 1.5){
  if(abs(x) <= k) x else sign(x) * k
}

A <- mean(unlist(lapply(geexex$Y1, function(y){
  x <- y - theta_cls
  -numDeriv::grad(psi_k, x = x)
})))

B <- mean(unlist(lapply(geexex$Y1, function(y){
  x <- y - theta_cls
  psi_k(x = x)^2
})))

## closed form covariance
Sigma_cls <- matrix(1/A * B * 1/A / nrow(geexex))
```

```
## $geex
## $geex$parameters
## [1] 4.82061
##
## $geex$vcov
##            [,1]
## [1,] 0.08356179
##
##
## $cls
## $cls$parameters
## [1] 4.999386
##
## $cls$vcov
##           [,1]
## [1,] 0.0928935
```

3.7.4 Example 7: Sample quantiles (approximation of ψ)

Approximation of ψ with geex is EXPERIMENTAL. Example 7 illustrates calculation of sample quantiles using M-estimation and approximation of the ψ function.

$$\psi(Y_{1i}, \theta) = \bigg(0.5 - I(Y_{1i} \leq \theta_1)\bigg)$$

```
SB7_estFUN <- function(data){
  Y1 <- data$Y1
  function(theta){
    0.5  - (Y1 <= theta[1])
  }
}
```

Note that ψ is not differentiable; however, `geex` includes the ability to approximate the ψ function via an `approx_control` object. The `FUN` in an `approx_control` must be a function that takes in the ψ function, modifies it, and returns a function of `theta`. For this example, I approximate ψ with a spline function. The `eval_theta` argument is used to modify the basis of the spline.

```
spline_approx <- function(psi, eval_theta){
  y <- Vectorize(psi)(eval_theta)
  f <- splinefun(x = eval_theta, y = y)
  function(theta) f(theta)
}
```

```
estimates <- m_estimate(
  estFUN = SB7_estFUN,
  data  = geexex,
  root_control = setup_root_control(start = 4.7),
  approx_control  = setup_approx_control(
    FUN = spline_approx,
    eval_theta = seq(3, 6, by = .05)))
```

A comparison of the variance is not obvious, so no comparison is made.

```
## $geex
## $geex$parameters
## [1] 4.7
##
## $geex$vcov
##           [,1]
## [1,] 0.1773569
##
##
## $cls
## $cls$parameters
## [1] 4.708489
##
## $cls$vcov
## [1] NA
```

3.7.5 Example 8: Robust regression

Example 8 demonstrates robust regression for estimating β from 100 observations generated from $Y_4 = 0.1 + 0.1X_{1i} + 0.5X_{2i} + \epsilon_i$, where $\epsilon_i \sim N(0,1)$, X_1 is defined as above, and the first half of the observation have $X_{2i} = 1$ and the others have $X_{2i} = 0$.

$$\psi_k(Y_{4i}, \theta) = \begin{pmatrix} \psi_k(Y_{4i} - \mathbf{x}_i^T \beta)\mathbf{x}_i \end{pmatrix}$$

```
psi_k <- function(x, k = 1.345){
  if(abs(x) <= k) x else sign(x) * k
}

SB8_estFUN <- function(data){
  Yi <- data$Y4
  xi <- model.matrix(Y4 ~ X1 + X2, data = data)
  function(theta){
    r <- Yi - xi %*% theta
    c(psi_k(r) %*% xi)
  }
}


estimates <- m_estimate(
  estFUN = SB8_estFUN,
  data  = geexex,
  root_control = setup_root_control(start = c(0, 0, 0)))
```

```
m <- MASS::rlm(Y4 ~ X1 + X2, data = geexex, method = 'M')
theta_cls <- coef(m)
Sigma_cls <- vcov(m)
```

```
## $geex
## $geex$parameters
## [1] -0.04050369  0.14530196  0.30181589
## 
## $geex$vcov
##             [,1]          [,2]          [,3]
## [1,]  0.05871932 -0.0101399730 -0.0133230841
## [2,] -0.01013997  0.0021854268  0.0003386202
## [3,] -0.01332308  0.0003386202  0.0447117633
## 
## 
## $cls
## $cls$parameters
## (Intercept)          X1          X2 
##  -0.0377206   0.1441181   0.2988842 
## 
## $cls$vcov
##             (Intercept)            X1            X2
## (Intercept)  0.07309914 -0.0103060747 -0.0241724792
## X1          -0.01030607  0.0020579145  0.0005364106
## X2          -0.02417248  0.0005364106  0.0431120686
```

3.7.6 Example 9: Generalized linear models

Example 9 illustrates estimation of a generalized linear model.

$$\psi(Y_i, \theta) = \left(D_i(\beta) \frac{Y_i - \mu_i(\beta)}{V_i(\beta)\tau} \right)$$

```
SB9_estFUN <- function(data){
  Y <- data$Y5
  X <- model.matrix(Y5 ~ X1 + X2, data = data, drop = FALSE)
  function(theta){
    lp <- X %*% theta
    mu <- plogis(lp)
    D  <- t(X) %*% dlogis(lp)
    V  <- mu * (1 - mu)
    D %*% solve(V) %*% (Y - mu)
  }
}
```

```
estimates <- m_estimate(
  estFUN = SB9_estFUN,
  data  = geexex,
  root_control = setup_root_control(start = c(.1, .1, .5)))
```

Compare point estimates to `glm` coefficients and covariance matrix to `sandwich`.

```
m9 <- glm(Y5 ~ X1 + X2, data = geexex,
          family = binomial(link = 'logit'))
theta_cls <- coef(m9)
Sigma_cls <- sandwich::sandwich(m9)
```

```
## $geex
## $geex$parameters
## [1] -1.1256071  0.3410619 -0.1148368
##
## $geex$vcov
##             [,1]         [,2]         [,3]
## [1,]  0.35202094 -0.058906883 -0.101528787
## [2,] -0.05890688  0.012842435  0.004357355
## [3,] -0.10152879  0.004357355  0.185455144
##
##
## $cls
## $cls$parameters
## (Intercept)          X1          X2
##  -1.1256070   0.3410619  -0.1148368
##
## $cls$vcov
##             (Intercept)           X1           X2
## (Intercept)  0.35201039 -0.058903546 -0.101534539
## X1          -0.05890355  0.012841392  0.004358926
## X2          -0.10153454  0.004358926  0.185456314
```

3.7.7 Example 10: Testing equality of success probabilities

Example 10 illustrates testing equality of success probablities.

$$\psi(Y_i, n_i, \theta) = \begin{pmatrix} \frac{(Y_i - n_i\theta_2)^2}{n_i\theta_2(1-\theta_2)} - \theta_1 \\ Y_i - n_i\theta_2 \end{pmatrix}$$

```
SB10_estFUN <- function(data){
  Y <- data$ft_made
  n <- data$ft_attp
  function(theta){
    p <- theta[2]
    c(((Y - (n * p))^2)/(n * p * (1 - p))  - theta[1],
      Y - n * p)
  }
}
```

```
estimates <- m_estimate(
  estFUN = SB10_estFUN,
  data  = shaq,
  units = 'game',
  root_control = setup_root_control(start = c(.5, .5)))
```

```
V11 <- function(p) {
  k    <- nrow(shaq)
  sumn <- sum(shaq$ft_attp)
  sumn_inv <- sum(1/shaq$ft_attp)
  term2_n  <- 1 - (6 * p) + (6 * p^2)
  term2_d <- p * (1 - p)
  term2  <- term2_n/term2_d
  term3  <- ((1 - (2 * p))^2) / ((sumn/k) * p * (1 - p))
  2 + (term2 * (1/k) * sumn_inv) - term3
}

p_tilde <- sum(shaq$ft_made)/sum(shaq$ft_attp)
V11_hat <- V11(p_tilde)/23

# Compare variance estimates
V11_hat
```

```
## [1] 0.0783097
```

```
vcov(estimates)[1, 1]
```

```
## [1] 0.1929791
```

```
# Note the differences in the p-values
pnorm(35.51/23, mean  = 1, sd = sqrt(V11_hat), lower.tail = FALSE)
```

```
## [1] 0.02596785
```

```
pnorm(coef(estimates)[1],
      mean = 1,
      sd = sqrt(vcov(estimates)[1, 1]),
      lower.tail = FALSE)
```

```
## [1] 0.1078138
```

This example shows that the empircal sandwich variance estimator may be different from other sandwich variance estimators that make assumptions about the structure of the A and B matrices.

CHAPTER 4: CAUSAL INFERENCE IN THE STUDY OF INFECTIOUS DISEASE

4.1 Introduction

The study of infectious disease and of causal inference have had a reciprocating influence over the past quarter century. The nature of infectious diseases motivated several important advances in causal inference theory and methods, from which the science of infectious disease has gained practical analytical tools. In studies of HIV treatments, for example, prescription and dosing of a drug may (i) depend on past treatments and biomarker values and (ii) in turn affect subsequent biomarker values. Without accounting for this dynamic evolution of a subject's treatment history, estimates of treatment efficacy may be biased. This motivated Robins and colleagues (Robins and Hernán, 2009) to develop the *g*-methods for estimating causal effects in the presence of time-varying confounding. Likewise, the importance of identifying immune surrogates of protection in vaccine trials led Gilbert and colleagues (Gilbert and Hudgens, 2008; Gilbert et al., 2014) to develop causal methods for evaluating potential surrogate biomarkers. Within the vaccine context, the need to quantify herd immunity effects led Halloran and collaborators (Halloran et al., 1991; Halloran and Struchiner, 1995; Hudgens and Halloran, 2008; Tchetgen Tchetgen and VanderWeele, 2012) to develop causal methods in the presence of interference.

In this chapter, we review how the potential outcomes approach to causal inference has contributed to the study of treatments and prevention of infectious diseases, as well as how infectious diseases accelerated causal inference research. Many of the methods presented here have been effective in answering important medical and public health questions about infectious disease. A motivation, a brief description, and applications are presented for each topic.

This chapter is organized as follows. After a review of elementary foundations and causal assumptions, a hypothetical, single-time point study introduces basic concepts fundamental to more complicated analyses. A more thorough introduction may be found in books such as Hernán and Robins (2017) or Morgan and Winship (2014). We then define time-varying confounding and present the *g*-methods as tools for handling such confounding. Next, three specific approaches are presented that address the crucial underlying assumption of no unmeasured confounding in different ways: test negative studies, negative controls, and regression discontinuity. We then introduce principal stratification and its application in vaccine studies. In all of those sections, the methods assume that the treatment of one individual does not affect outcomes of other individuals, a phenomenon known as no interference (Cox, 1958). Due to the interdependent nature of many infectious diseases, interference has been the subject of much causal inference research in the past ten years. In the last section, some advances in causal inference in the presence of interference are discussed.

Throughout, the following general notation is used. Let Y_i be an outcome of interest, such as infection status or CD4 count, for individual $i = 1, \ldots, n$ in a study. Let A_i be an exposure, treatment, or intervention of interest, such as vaccination, drug treatment, or hand washing. Observed values are denoted in lower case, e.g., y_i, a_i. Except where explicit, A_i is considered binary, so that subjects are either exposed (1) or unexposed (0). Boldface represents vectors (or matrices where appropriate) of variables. For example, $\mathbf{A}$ is the vector of all subjects' exposures, i.e., $\mathbf{A} = (A_1, A_2, \ldots, A_n)$. Measured covariates, denoted L_i, may influence both the assignment to intervention as well as the outcome. This notation is modified as needed.

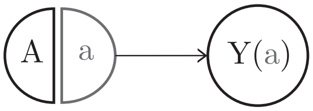
Many philosophies and frameworks exist for assessing causality. This chapter and the methods presented herein are couched in the potential outcomes framework, dating back to 1923 in the statistics literature (Splawa-Neyman et al., 1990).

4.2 Causal assumptions

Causal inference using potential outcomes (hereafter causal inference) begins with the causal estimand. The estimand makes explicit how potential outcomes may vary depending on a treatment assignment. Many causal applications invoke the *stable unit treatment value assumption* (SUTVA) (Rubin, 2005), which includes an assumption of no interference. The assumption supposes that a subject's potential outcomes do not depend on the treatment of other subjects. When treatment is binary, SUTVA implies that subjects have only two potential outcomes, denoted by $Y(0)$ and $Y(1)$. The fundamental problem of causal inference (Holland, 1986) is that typically only one potential outcome for a subject can be observed in a study; thus individual-level effects cannot be identified. Population-level estimands, though, may be identified under certain assumptions, and this summary of individual-level potential outcomes is chosen as the target of inference based on the research question(s). The average causal effect ($E[Y(1)-Y(0)]$), for example, is a common estimand in randomized controlled trials.

Many of the approaches we introduce below are illustrated using causal graphs, which can help in establishing the assumptions necessary for identification of an estimand. Graphical tools such as directed acyclic graphs (DAGs) (Pearl, 2010a) are useful in visualizing and mapping relationships between the covariates, intervention(s), and outcomes. DAGs are well-developed tools for reasoning about causal relationships (Pearl, 2009; Joffe et al., 2012). In a causal DAG, nodes represent variables and edges represent causal relations between variables. DAGs, however, typically do not make potential outcomes explicit. Richardson and Robins (2013) unify the potential outcomes and graphical approaches to causal inference through single world intervention graphs (SWIGs) and single world intervention templates (SWITs). Constructing a SWIG or SWIT from a DAG entails splitting treatment nodes into a random part A and a fixed part a, and then replacing descendants of the treatment

Figure 4.14: A denotes random treatment assignment, a denotes the treatment of interest, and $Y(a)$ denotes the potential outcome corresponding to a.



nodes with potential outcomes for treatment a. We use SWITs to clarify the methods in this chapter.

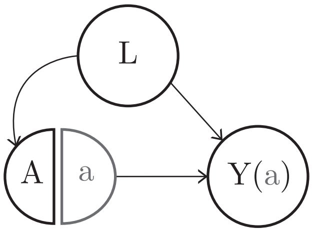
Consider a simple experiment in which all individuals (or units) are randomized to receive treatment or not with equal probability such that $Y(a) \perp A$ for all a where $\perp$ indicates independence. Figure 4.14 illustrates this simple experiment in a SWIT. There are no arrows from A to $Y(a)$, hence the treatment variable is independent of the potential outcome. In a similar way, if randomization is conditional on measured covariates, one may assume conditional exchangeability, i.e.,

$$Y(a) \perp A|L = l \text{ for all } a \text{ and } l. \tag{4.A1}$$

Consider a randomized experiment with two strata defined by a single covariate. Subjects in one stratum are randomized to treatment with probability p, and subjects in the other stratum are randomized with probability $p' \neq p$. If stratum membership is associated with the potential outcomes, then unconditional exchangeability $Y(a) \perp A$ will not hold in general. On the other hand, conditional exchangeability (4.A1) will be satisfied in such a randomized experiment. Conditional exchangeability is also known as ignorability or the no unmeasured confounding assumption (Hernán and Robins, 2017; Pearl, 2009; Richardson and Robins, 2013).

A defining characteristic of observational studies is that the assignment (or selection) mechanism is not known. Hence, assumptions are often made about the assignment mechanism in order to draw causal inferences in the observational setting. A typical assumption

Figure 4.15: The covariate(s) L confounds the relationship between A and $Y(a)$. When L contains all variables that block the paths between A and $Y(a)$, conditional exchangeability holds.



asserts that given certain baseline covariates L, conditional exchangeability holds. The assumption must be based on scientific knowledge in an observational setting. Figure 4.15 presents a SWIT for both the conditionally randomized experiment and an observational study under assumption (4.A1). When (4.A1) holds, the only path from $Y(a)$ to A is through L.

If all units receive treatment by design, the study data contain no information about the control group. Clearly, any treatment versus control causal effects cannot be identified from such a study. The assumption of positivity formalizes this notion. To identify causal effects pertaining to a certain level of a, there must be some probability of receiving a. In observational studies where conditional exchangeability is assumed, identifiability also requires:

$$\Pr(A = a|L = l) > 0 \text{ for all } l \text{ where } \Pr(L = l) \neq 0. \tag{4.A2}$$

Assumptions (4.A1) and (4.A2) are generally the basis for identification of causal effects of a point exposure. For further details, Hernán and Robins (2006) and Hernán and Robins (2017, ch. 3) provide excellent introductions on the fundamental assumptions of causal inference.

4.3 Causal inference for single and multiple point exposures

The crux of identifying and estimating causal effects is expressing potential outcomes as functions of observable random variables. For example, a population mean potential outcome is often a target estimand, for which standardization or inverse probability weighting (IPW) are typical estimation methods (Hernán and Robins, 2006). Standardization averages across the population strata of L. In particular, for discrete L under assumptions (4.A1) and (4.A2), it follows that

$$E[Y(a)] = \sum_{l} E[Y|A = a, L = l] \Pr(L = l), \tag{4.10}$$

where the summation is over all levels of l. The right hand side of this equation is a function of observable random variables only and thus is composed of identifiable parameters. Similarly, under (4.A1) and (4.A2), it follows that

$$E[Y(a)] = E[I(A = a)Y/ \Pr(A|L)] \tag{4.11}$$

where the right side of (4.11) is also identifiable. Expressions (4.10) and (4.11) suggest two estimators of $E[Y(a)]$ by replacing identifiable parameters with consistent or unbiased estimators. In cases where L is discrete and low-dimensional, non-parametric estimators may be used, in which case, the IPW and standardization estimators are equal. In more complicated situations, when L contains multiple variables for example, parsimonious low-dimensional models are often assumed to deal with the curse of dimensionality (Hernán and Robins, 2017, ch. 10). In this case, the standardization and IPW estimators will not necessarily be equal.

4.3.1 Time-varying exposures and the *g*-methods

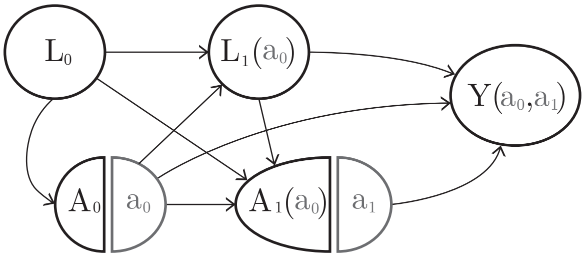
Infectious disease interventions may involve treatments at multiple points in time. Hernán et al. (2002) estimate a causal effect of the drug zidovudine on HIV patients' CD4 count. Over 16 visits, a patient's treatment history was recorded as $\overline{A} = [A(1), \ldots, A(16)]$, where $A(t) = 1$ indicates a subject was taking zidovudine at time (visit) t. Similarly, $\overline{L}$ records a patient's covariate history. Let $\overline{a}_t$ be a fixed treatment history up to time t. An important policy question is which treatment regimes maximize potential CD4 counts. To answer this, Hernán et al. (2002) target the estimand, $E[Y_{t+1}(\overline{a}_t) - Y_{t+1}(\overline{0}_t)]$, which is the average difference in potential outcomes at visit $t+1$ under treatment regime $\overline{a}_t$ versus no treatment at any point ($\overline{0}_t$). Inference about this estimand in observational studies is challenging, especially if certain time-varying covariates $L(t)$, such as viral load, are affected by prior treatment and affect subsequent treatment and the outcome.

Marginal structural models, the *g*-formula, and structural nested models were developed by Robins and others (Robins, 1989; Robins and Rotnitzky, 1992; Robins, 1993) to account for confounding by variables affected by the exposure itself as shown in Figure 4.16. In this graph, the covariate(s) L has three features: (1) it is associated with $Y(a)$, (2) it is affected by exposure A_0, and (3) it predicts exposure A_1 (Vansteelandt and Joffe, 2014). Without accounting for this structure, estimates of the effect of (A_0, A_1) on Y may be biased. Robins (1997; see also Robins and Hernán, 2009) details the sequential ignorability assumptions and methods for estimating parameters from these models: inverse probability weighted estimators for marginal structural models, G-computation for the *g*-formula, and G-estimation for structural nested models. We begin with marginal structural models, which have become popular, partly due to ease of implementation.

4.3.1.1 Marginal structural models

A marginal structural model (MSM) entails modeling some aspect of the distribution of $Y(a)$, typically the mean, as a function of the exposure. For example, one parameterization

Figure 4.16: Example of time-varying confounder (L_1) affected by prior treatment.



for the MSM for the graph in Figure 4.16 is $E[Y(\mathbf{a})] = \beta_0 + \beta_1 a_0 + \beta_2 a_1 + \beta_3 a_0 a_1$. The choice of the MSM is driven by the scientific or policy questions. Inverse probability of treatment weighted (IPTW) estimators are used to estimate parameters from an MSM. Details on estimating parameters from MSMs can be found in Robins et al. (2000).

Sterne et al. (2005) used marginal structural Cox models to estimate the effect of highly active antiretroviral therapy (HAART) from a large observational cohort study of HIV positive individuals. Their results suggest a simple Cox model that does not account for time-dependent confounding underestimates the positive effect of HAART. The hazard ratios based on the MSM showed HAART therapy has a statistically significant effect, while methods that do not account for the time-dependent confounding did not. This highlights the need to account for time-varying confounding appropriately.

4.3.1.2 Parametric *g*-formula

The g-formula generalizes the notion of standardization where the expectation of the outcome given treatment and covariates is weighted across the conditional distribution(s) of the confounding covariates:

$$E[Y(\overline{a})] = \int_{\overline{l}} E[Y|\overline{A} = \overline{a}, \overline{L} = \overline{l}] \prod_{t=1}^{T} dF(l_t|\overline{A}_{t-1} = \overline{a}_{t-1}, \overline{L}_{t-1}) \tag{4.12}$$

where $dF(l_t|\overline{A}_{t-1} = \overline{a}_t, \overline{L}_{t-1})$ denotes the conditional distribution of L_t given the exposure and covariate histories, up to time $t-1$. In longitudinal observational studies with multiple confounders, the form of $E[Y|\overline{A} = \overline{a}, \overline{L} = \overline{l}]$ and $dF_{\overline{l}_t|\overline{A}_{t-1},\overline{L}_{t-1}}$ are unknown. The parametric g-formula plugs in parametric models for the conditional mean outcome ($E[Y|\overline{A} = \overline{a}, \overline{L} = \overline{l}]$) and conditional distribution of the covariates ($dF_{\overline{l}_t|\overline{A}_{t-1},\overline{L}_{t-1}}$). Multiple time points and multiple covariates in $\overline{L}$ compound the analyst's work. A model is specified for covariates at each time point conditional on past exposure and past covariates in addition to a mean outcome model. Importantly, certain parameterizations of $E[Y|\overline{A} = \overline{a}, \overline{L} = \overline{l}]$ and $dF(l_t|\overline{A}_{t-1} = \overline{a}_t, \overline{L}_{t-1})$ result in the g-null paradox (Robins and Hernán, 2009), wherein the sharp null hypothesis of no treatment effect will be falsely rejected in large samples regardless of the data. When the conditional mean of Y given $\overline{A}$ and $\overline{L}$ and the conditional distribution of L_t given $\overline{A}_{t-1}$ and $\overline{L}_{t-1}$ are estimated nonparametrically, the g-formula does not suffer from the g-null paradox.

Westreich et al. (2012) and Schomaker et al. (2013) both use G-computation of the parametric g-formula to study HIV antiretroviral therapy in different cohorts. Their papers provide detailed step-by-step instructions for implementing the G-computation algorithm, which uses Monte Carlo simulations to approximate the integral in (4.12). Briefly, the steps involve fitting models for the outcome and covariates followed by a process of simulating data from these models for each treatment regime of interest, estimating treatment effects from the simulated data, and bootstrapping to obtain confidence intervals.

4.3.1.3 Structural nested models

G-estimation of structural nested models (SNMs) is perhaps the least used of the three models, partly owing to lack of standard software for fitting these models. But they also have certain advantages (Vansteelandt and Joffe, 2014). For a point treatment, a structural mean model directly models a causal effect, generally the effect of removing treatment: $h(E[Y(a)|L = l, A = a]) - h(E[Y(0)|L = l, A = a]) = \psi(a, l; \beta)$, where h is some link function such as the identity, log, or logit. The function $\psi(a, l; \beta)$ parameterizes the causal contrast such that $\psi(0, l; \beta) = 0$ for all l and β. For instance, with h as the identity function, an additive SNM is $\psi(a, l; \beta) = (\beta_0 + \beta_1 l)a$. The causal parameters β can be identified from this model with a weaker version of (4.A1). One only needs to assume that $Y(0) \perp A|L$.

The "nested" in SNM refers to the process by which the effect of treatment is removed one time point at a time. Vansteelandt and Joffe (2014) give an overview of the scope and application of SNMs for estimating joint exposure effects from multiple time points. MSMs have some disadvantages compared to SNMs (Robins and Hernán, 2009). For example, IPW estimators of MSM parameters can be highly variable, leading to imprecise inferences. SNMs allow direct modeling of effect modification by time-dependent covariates and can also be used to estimate causal effect when treatment is confounded but an instrumental variable is available. Sensitivity analysis models for SNMs tend to be less restrictive and more useful than those for MSMs.

Garn et al. (2016) fit a SNM using G-estimation to draw inference about the effect of school water and sanitation improvements on student health outcomes, particularly diarrhea and helminth infections. Previous studies had estimated intent-to-treat effects of the public health interventions from a cluster-randomized trial of 185 schools in Kenya. Garn et al. (2016) used the randomization assignment as an instrumental variable and then accounted for actual adherence of the schools to the interventions using SNMs.

4.4 Alternative approaches to address confounding

The methods presented above describe causal inference *de rigueur* with either point or longitudinal exposures. This section presents alternative approaches to address confounding in observational studies motivated by particular applications in infectious disease. Test negative studies, whose theoretical causal properties were only recently examined, are a cost-effective design for measuring vaccine effectiveness from routine surveillance data. Negative controls can, in some cases, provide a tool for assessing the no unmeasured confounding assumption (4.A1). Lastly, regression discontinuity is a means to perform causal inference when treatment decisions are made based on a threshold on a continuous variable.

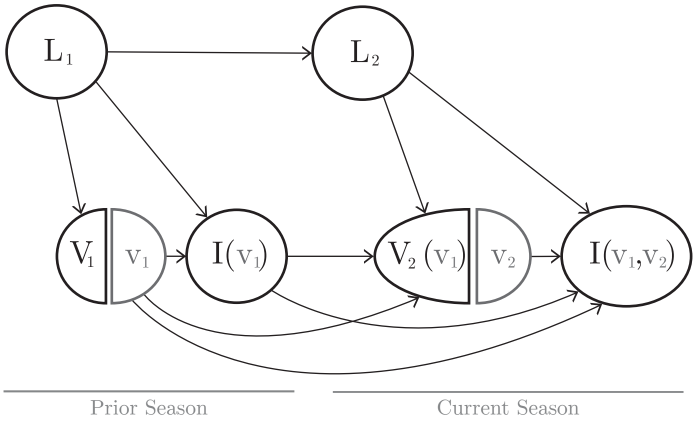
4.4.1 Test negative design

Influenza vaccine effectiveness (VE) varies from year to year due to the frequent antigenic changes in influenza viruses. A popular study design for measuring VE in such a dynamic environment is the test negative design (Jackson and Nelson, 2013). Patients presenting to health clinics with influenza-like symptoms are tested for influenza viruses. Patient vaccination history and other covariate histories are also recorded. Those who test positive are classified as cases, and those who test negative are non-cases. The odds ratio comparing the odds of being a case in vaccinated versus unvaccinated subjects from logistic regression provides an estimate for VE, after controlling for confounding. This study design is relatively inexpensive and easy to implement.

The test negative design is meant to reduce confounding bias from health-seeking behavior and case misclassification. Sullivan et al. (2016) use causal diagrams to explain possible biases that may arise from the design. Two sources of confounding in influenza vaccine studies they considered are prior exposure/vaccination history and health-seeking behavior. Prior exposure history could be an important source of confounding in influenza vaccine studies (Figure 4.17), but the degree to which this confounding can be addressed depends on whether prior exposures are measured and appropriately controlled for. Furthermore, test negative designs do not inherently remove bias due to health-seeking behavior. In including subjects tested for influenza, the test negative design may induce selection bias as influenza testing is affected both by infection and health-seeking behavior. Sullivan et al. (2016) conclude that the design, as in any observational study, may reduce but cannot eliminate confounding or selection bias.

Schwartz et al. (2017) compared rotavirus VE estimates from a test-negative design to VE estimates from two randomized controlled trials. They found, as others have found for influenza (De Serres et al., 2013), the estimates to be nearly equal.

Figure 4.17: Graph representing a test negative design. V_t and I_t represent vaccination status and infectious status, respectively, at time t. Prior infection status may confound the relationship between current year vaccination and infection status.

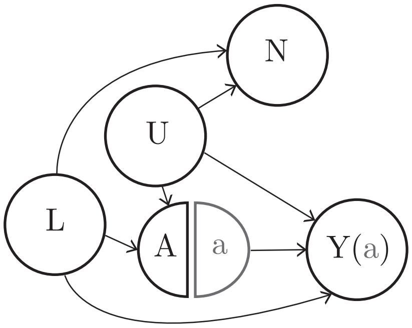


4.4.2 Negative controls

Lipsitch et al. (2010) describe the use of negative controls – common in experimental bench science – for epidemiology. Negative control outcomes or negative control exposures may detect residual confounding. Lipsitch et al. (2010) give several examples including Jackson et al. (2006) who found that influenza VE in seniors had been overestimated. Because an influenza vaccine can only reduce influenza-related morbidity or mortality by preventing influenza, they argue that a causal protective effect would be strongest during influenza season. They used pneumonia/influenza hospitalization after vaccination but before influenza season as the negative control. If a vaccination effect is observed after vaccination season (fall) but before the influenza season (winter), they reasoned, the vaccine's apparent effect may be due to confounding. In their analysis, Jackson et al. (2006) estimated that the protective effect was greatest before the influenza season, and thus concluded that residual confounding accounts for a portion of the vaccine's estimated effect among seniors.

Another example presented by Lipsitch et al. (2010) with application to infectious diseases reveals the utility of negative controls in detecting recall bias. Zaadstra et al. (2008)

Figure 4.18: N meets the criteria to serve as a negative control outcome to check for residual confounding. $Y(a)$ and N share all the same parents except a. An association between A and N, given L, suggests residual confounding may be present.



studied whether particular childhood infections are associated with multiple sclerosis (MS). MS cases and controls were asked about childhood infections. Because MS cases may have greater and more specific recall, study participants were also asked negative control exposure questions about childhood medical events not plausibly linked to MS.

Lipsitch et al. (2010) graphically represent the ideal negative control outcome or negative control exposure. A negative control outcome N should satisfy the criterion that N and $Y(a)$ share all the same incoming arrows from measured confounders L and unmeasured confounders U *except* for a (Figure 4.18). A non-null association between A and N suggests uncontrolled confounding. But the magnitude and form of the bias cannot be determined from this finding.

Weisskopf et al. (2016) consider whether an imperfect negative control, where an association exists between a negative control exposure, B, and Y, could affect results. They conclude that, under certain assumptions, negative control exposures can still be useful for detecting residual confounding even if one is unsure of a causal relation between B and Y. Fisher (2016) studied the untestable assumptions necessary to apply negative controls. Using the approach of Richardson et al. (2015) to adjust standardized mortality ratios for uncontrolled confounding, she showed that the choice of the negative control outcome can influence results. In some cases, adjusting for the negative control outcome can increase bias due to residual confounding rather than decrease it. Thus presenting both adjusted and unadjusted estimates is recommended in practice.

4.4.3 Regression Discontinuity

When a binary treatment is assigned based on a threshold rule on an underlying continuous variable, the methods of regression discontinuity may be applicable. Heuristically, regression discontinuity methods rely on the idea that in the region around the treatment threshold, say t, of the continuous (or "running") variable Z, the treatment assignment can be considered as if approximately random (Bor et al., 2014; Cattaneo et al., 2015). Regression discontinuity has origins in econometrics, but Bor et al. (2014) describe how the method may have broad utility in epidemiology. Bor et al. (2014) give the example of prescribing antiretroviral therapy to HIV infected individuals based on a CD4 threshold value. CD4 counts are used to determine eligibility for antiretroviral therapy. If a cutoff of, say, 200 cells/μL of blood was used by doctors to assign therapy, a regression discontinuity design could be appropriate.

The threshold mechanism may not be deterministic or "strict." "Fuzzy" regression discontinuity designs account for a probabilistic mechanism where the intervention is assigned based on the threshold t plus other factors such as clinical judgment (Bor et al., 2014). For example, a physician may prescribe antiretroviral therapy to a patient with CD4 count greater than 200 based on other considerations, such as the patient's viral load. Bor et al. (2015) consider three levels of assumptions regarding how conditional regression discontinuity estimands, $E[Y(a)|Z]$, vary over Z. The most restrictive approach makes a strong assumption about the functional form of $E[Y(a)|Z]$ and allows for extrapolation across the range of Z. If the assumed functional form is wrong, estimates may be severely biased. Another approach makes a continuity assumption at the cutoff and allows for extrapolation in the region of the cutoff. The third approach assumes that patients with Z

values close to the cutoff are truly randomly assigned to A and the data essentially represent a randomized trial in the region of the cutoff.

4.5 Principal stratification

Principal stratification was originally motivated by treatment comparisons that require adjusting for or conditioning on post-treatment variables, such as adjusting for noncompliance in clinical trials (Frangakis and Rubin, 2002). Using this approach, units are grouped into basic principal strata defined by the joint potential values of the post-treatment variable(s) under each of the treatments being compared. Then, principal strata are formed by unions of the basic principal strata. Causal estimands are defined by comparing potential outcomes within a principal stratum. The key property of principal strata is they are not affected by treatment assignment and therefore conditioning on a principal strata is not subject to selection bias. For discussion of the strengths and drawbacks of principal stratification, see Pearl (2011), Gilbert et al. (2011), Joffe (2011), and VanderWeele (2011).

Two important applications of principal stratification arise in vaccine trials. First, in vaccine studies it may be of interest to assess the effect of vaccination on disease or other post-infection outcomes among the subset of individuals who become infected with the pathogen of interest during the study. For example, in HIV vaccine trials, it is of interest to compare viral load among individuals who become infected during the trial (Gilbert et al., 2003). In this case, infection status plays the role of the post-treatment (i.e., post-vaccination) variable. Second, determining an immunological surrogate of protection is an important component of HIV vaccine research. Here the post-treatment variable of interest is the immune response measured after vaccination. Below we consider the utility of the principal stratification framework in these two applications.

4.5.1 Post-infection selection

To assess the effect of vaccination on post-infection outcomes, it is natural to compare the distribution of the outcome between vaccinated individuals who become infected and unvaccinated individuals who become infected. Unfortunately, conditioning on infection may create selection bias. In particular, comparisons between vaccinated and unvaccinated individuals after conditioning on infection do not necessarily have causal interpretations without the strong assumption that vaccinated persons who become infected are exchangeable with unvaccinated persons who become infected.

Hudgens and Halloran (2006) give the example of a vaccine that protects individuals with strong immune systems, hence those who are vaccinated but become infected have weaker immune systems on average compared to infected controls. A naive comparison of infected vaccinated individuals and infected controls may lead to the incorrect conclusion that vaccination results in worse post-infection outcomes for the vaccinated. Using a principal stratification approach, Hudgens and Halloran (2006) develop estimands, identifiability assumptions, and maximum likelihood estimators for post-infection effects on binary outcomes of vaccinations (see also Gilbert et al., 2003; Hudgens et al., 2003).

Consider a hypothetical vaccination trial where subjects are randomly assigned to placebo $A = 0$ or vaccine $A = 1$. Let $S(a) = 1$ indicate post-vaccination infection and $S(a) = 0$ indicate no post-vaccination infection if an individual receives treatment a. Similarly, let $Y(a)$ denote a binary post-infection potential outcome for treatment a. Because S is measured after randomization, a comparison of the observed distributions $\Pr(Y = y|S = s, A = 0)$ and $\Pr(Y = y|S = s, A = 1)$ does not necessarily have a causal interpretation because the set of individuals who would have $S = s$ when $A = 0$ will not in general be equal to the set of individuals who would have $S = s$ when $A = 1$.

A basic principal stratification P_0 is a partition of the units into strata that have the same $\{S(0), S(1)\}$ vector. In this example, these strata are subjects who (i) would be uninfected with both vaccine and placebo $\{S(0) = 0, S(1) = 0\}$, (ii) would be uninfected after receiving placebo but infected with vaccination $\{0,1\}$, (iii) would be infected after receiving placebo but uninfected with vaccination $\{1,0\}$, and (iv) would be infected with either placebo or vaccine $\{1,1\}$. A principal stratification P is the union of any of the sets in P_0. The vaccine's effect on the post-infection outcome Y is then defined by some contrast in distribution of $Y(0)$ and $Y(1)$ within the $\{1,1\}$ principle strata. Such estimands quantify the vaccine effect on Y among individuals who will become infected regardless of whether they are vaccinated. In general, effects within principal strata are only partially identifiable, such that bounds or sensitivity analyses may be employed to draw inference (e.g., see Gilbert et al., 2003; Hudgens and Halloran, 2006). A harder problem is to estimate effects across principal strata, as the case with principal surrogates.

4.5.2 Principal Surrogates

Use of surrogate biomarkers can facilitate vaccine development and inform vaccine policy. A careful distinction must be made between a statistical surrogate and a principal surrogate, where the latter retains a causal interpretation. A principal surrogate is defined in terms of potential outcomes. Following Frangakis and Rubin (2002), Gilbert and Hudgens (2008) define S to be a principal surrogate for Y if A has an effect on S if and only if A has an effect on Y. Thus they require that $\Pr\{Y(0)|S(1) = S(0) = s\} = \Pr\{Y(1)|S(1) = S(0) = s\}$ for all s, and that $\Pr\{Y(0)|S(1) = s_1, S(0) = s_0\} \neq \Pr\{Y(1)|S(1) = s_1, S(0) = s_0\}$ for $s_1 \neq s_0$.

Gilbert and colleagues have developed methods for evaluating whether immune responses induced by vaccination are valid surrogates for a vaccine's effect on infection or disease (e.g., see Gilbert et al., 2014, 2016). Gilbert et al. (2014) used this approach to demonstrate that fold rise in antibody titers as measured by a certain assay is a principal surrogate of the effect of live, attenuated zoster vaccination on the incidence of herpes zoster. Gilbert et al. (2016) discuss use of the principal surrogates framework in the design of upcoming HIV vaccine trials.

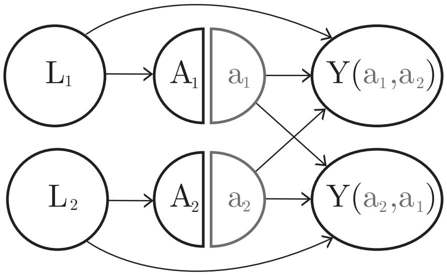
4.6 Interference

The causal estimands and associated inferential methods described thus far all assume no interference. This assumption presumes each subject's potential outcomes are not affected by the treatment of other subjects. The contagious nature of some infectious diseases invalidates this assumption. Halloran and Struchiner (1995) recognized this problem in infectious disease studies. Much progress has been made in the intervening years; see Halloran and Hudgens (2016) for a recent review.

Consider a study of $n = 3$ subjects and suppose we allow for the possibility there is interference between individuals. Then in the case of a binary treatment, each subject will have eight potential outcomes $Y(a_1, a_2, a_3)$ for each possible treatment vector $\{(0,0,0),(1,0,0),\ldots,(1,1,1)\}$. As n increases, the number of potential outcomes increases exponentially, making inference challenging. To reduce the number of potential outcomes, the partial interference assumption (Hudgens and Halloran, 2008; Sobel, 2006) is often used, where it is assumed that subjects can be partitioned into groups such that interference may occur within a group but not between groups.

Hudgens and Halloran (2008) defined four estimands under partial interference. The direct, or unit-level, effect measures the average effect of treatment on a subject while holding the treatment coverage (the proportion of others within the group who receive treatment) fixed. The indirect, or spillover, effect measures the difference in potential outcomes for different levels of treatment coverage among the untreated. Spillover effects can also be defined in treated individuals. The total effect is the sum of direct and indirect effects. The overall effect compares average potential outcomes under different treatment coverages. Hudgens and Halloran (2008) also proposed unbiased estimators for these four estimands for two-stage randomized trials. In the context of observational studies, Tchetgen Tchetgen and VanderWeele (2012) consider inference about these effects under a modification of the conditional exchangeability assumption (4.A1); namely, they assume

Figure 4.19: Each row represents a subject. Arrows crossing from the treatment level into the other subject's potential outcome denote interference.



an individual's potential outcomes are conditionally independent of the group's treatment vector, $Y_{ij}(\mathbf{a}_i) \perp \mathbf{A}_i | \mathbf{L}_i$, where i now indexes groups and j indexes individuals within a group.

The Centre for Diarrhoeal Disease Research conducted a cholera vaccine trial in the area of Matlab, Bangladesh (Ali et al., 2005). Ali et al. (2005) reported a non-causal analysis finding an association between the incidence of cholera in unvaccinated individuals and the level of vaccine coverage in their neighborhood, suggesting a herd immunity effect for the vaccine. Using estimators for observational data developed by Tchetgen Tchetgen and VanderWeele (2012), Perez-Heydrich et al. (2014) analyzed the same data and corroborated Ali et al. (2005) with a causal analysis.

Ogburn and VanderWeele (2014) considered causal graphs illustrating interference between subjects. Figure 4.19 shows one such graph for groups of two subjects with single point treatment. As the number of subjects and points in time increase, depicting the many types of interference becomes increasingly challenging.

Recent work has focused on relaxing the partial interference assumption and estimating causal effects in networks or groups of networks. For example, van der Laan (2014) considered semi-parametric inference about causal effects on a single network of individuals. Drawing inference based on a single realization of a network is challenging because standard large sample frequentist approaches generally require independent replicates. Athey et al.

(2016) considered an alternative perspective, deriving randomization-based tests of causal effects based on a single network of connected units.

A vaccine may have two types of indirect effects from vaccinated individuals. Contagion effects describe the effect of a vaccine preventing a subject from acquiring the target pathogen and thus transmitting it. Infectiousness effects describe the effect of vaccine in reducing a person's capacity to spread an infection, thus decreasing the transmission probability (VanderWeele et al., 2012). Ogburn and VanderWeele (2017) considered estimands and causal assumptions for contagiousness and infectiousness effects from social network data. They propose using modified generalized linear models to estimate these effects.

4.7 Summary

We reviewed several study designs and causal analysis methods targeted for application in the study of infectious diseases, such as the g-methods, regression discontinuity, negative controls, and test negative study design. We discussed how certain infectious disease problems motivated new developments in causal inference, such as principal stratification and relaxing the no interference assumption.

BIBLIOGRAPHY

Ahuja, S., editor (2013). *Monitoring Water Quality: Pollution Assessment, Analysis, and Remediation*. Elsevier, Amsterdam; Boston.

Ali, M., Emch, M., von Seidlein, L., Yunus, M., Sack, D. A., Rao, M., Holmgren, J., and Clemens, J. D. (2005). Herd immunity conferred by killed oral cholera vaccines in Bangladesh: a reanalysis. *The Lancet*, 366(9479):44–49.

Athey, S., Eckles, D., and Imbens, G. W. (2016). Exact p-values for network interference. *Journal of the American Statistical Association*, In press.

Boos, D. D. and Stefanski, L. A. (2013). *Essential Statistical Inference Theory and Methods*. Springer, New York, NY.

Bor, J., Moscoe, E., and Brnighausen, T. (2015). Three approaches to causal inference in regression discontinuity designs. *Epidemiology*, 26(2):e28–e30.

Bor, J., Moscoe, E., Mutevedzi, P., Newell, M.-L., and Brnighausen, T. (2014). Regression discontinuity designs in epidemiology. *Epidemiology*, 25(5):729–737.

Burkholder, J. M. (2002). *Cyanobacteria*, pages 952–982. Wiley, New York.

Cape Fear River Partnership (2013). Cape Fear River basin action plan for migratory fish. Technical report.

Carey, V. J. (2015). *gee: Generalized Estimation Equation Solver*. Ported to R by Thomas Lumley and Brian Ripley; R package version 4.13-19.

Cattaneo, M. D., Frandsen, B. R., and Titiunik, R. (2015). Randomization inference in the regression discontinuity design: An application to party advantages in the U.S. Senate. *Journal of Causal Inference*, 3(1):1–24.

Cox, D. R. (1958). *Planning of Experiments*. Wiley, New York.

Dame, R., Alber, M., Allen, D., Mallin, M., Montague, C., Lewitus, A., Chalmers, A., Gardner, R., Gilman, C., and Kjerfve, B. (2000). Estuaries of the South Atlantic coast of North America: their geographical signatures. *Estuaries*, 23(6):793–819.

De Serres, G., Skowronski, D. M., Wu, X. W., and Ambrose, C. S. (2013). The test-negative design: validity, accuracy and precision of vaccine efficacy estimates compared to the gold standard of randomised placebo-controlled clinical trials. *Eurosurveillance*, 18(37):2–10.

Di Gennaro, D. and Pellegrini, G. (2016). Policy evaluation in presence of interferences: A spatial multilevel DID approach. Technical report, CREI Working Paper No. 4/2016. Centro di Ricerca Interdipartimentale di Economia delle Istituzioni (CREI), University of Rome.

Dodds, W. K., Bouska, W. W., Eitzmann, J. L., Pilger, T. J., Pitts, K. L., Riley, A. J., Schloesser, J. T., and Thornbrugh, D. J. (2009). Eutrophication of U.S. freshwaters: Analysis of potential economic damages. *Environmental Science & Technology*, 43(1):12–19.

Dubbs, L. L. and Whalen, S. C. (2008). Light-nutrient influences on biomass, photosynthetic potential and composition of suspended algal assemblages in the middle Cape Fear River , USA. *International Review of Hydrobiology*, 93(6):711–730.

Duembgen, L., Nordhausen, K., and Schuhmacher, H. (2014). *fastM: Fast Computation of Multivariate M-estimators*. R package version 0.0-2.

EPA, U. (2015). A compilation of cost data associated with the impacts and control of nutrient pollution. Technical report, U.S. Environmental Protection Agency.

Fay, M. P. and Graubard, B. I. (2001). Small-sample adjustments for Wald-type tests using sandwich estimators. *Biometrics*, 57(4):1198–1206.

Fisher, L. H. (2016). *Modeling of Infectious Disease Surveillance Data*. PhD thesis, University of Washington.

Frangakis, C. E. and Rubin, D. B. (2002). Principal stratification in causal inference. *Biometrics*, 58(1):21–29.

Freedman, D. A. (2005). Statistical models for causation.

Garn, J. V., Brumback, B. A., Drews-Botsch, C. D., Lash, T. L., Kramer, M. R., and Freeman, M. C. (2016). Estimating the effect of school water, sanitation, and hygiene improvements on pupil health outcomes. *Epidemiology*, 27:752.

Gilbert, P. B., Bosch, R. J., and Hudgens, M. G. (2003). Sensitivity analysis for the assessment of causal vaccine effects on viral load in HIV vaccine trials. *Biometrics*, 59(3):531–541.

Gilbert, P. B., Gabriel, E. E., Miao, X., Li, X., Su, S.-C. C., Parrino, J., and Chan, I. S. F. (2014). Fold rise in antibody titers by measured by glycoprotein-based enzyme-linked immunosorbent assay is an excellent correlate of protection for a herpes zoster vaccine, demonstrated via the vaccine efficacy curve. *Journal of Infectious Diseases*, 210(10):1573–81.

Gilbert, P. B., Huang, Y., and Janes, H. E. (2016). Modeling HIV vaccine trials of the future. *Current Opinion in HIV and AIDS*, 11(6):620–627.

Gilbert, P. B. and Hudgens, M. G. (2008). Evaluating candidate principal surrogate endpoints. *Biometrics*, 64(4):1146–1154.

Gilbert, P. B., Hudgens, M. G., and Wolfson, J. (2011). Commentary on 'Principal stratification–a goal or a tool?' by Judea Pearl. *International Journal of Biostatistics*, 7(1):1–15.

Halekoh, U., Hojsgaard, S., and Yan, J. (2006). The R package geepack for generalized estimating equations. *Journal of Statistical Software*, 15(2):1–11.

Halloran, M. E., Haber, M., Longini, I. M., and Struchiner, C. J. (1991). Direct and indirect effects in vaccine efficacy and effectiveness. *American Journal of Epidemiology*, 133(4):323–331.

Halloran, M. E. and Hudgens, M. G. (2016). Dependent happenings: a recent methodological review. *Current Epidemiology Reports*, 3(4):297–305.

Halloran, M. E. and Struchiner, C. J. (1995). Causal inference in infectious diseases. *Epidemiology*, 6(2):142–151.

Hennig, C. (2012). *smoothmest: Smoothed M-estimators for 1-dimensional location*. R package version 0.1-2.

Hernán, M. A., Brumback, B., and Robins, J. M. (2000). Marginal structural models to estimate the causal effect of zidovudine on the survival of HIV-positive men. *Epidemiology*, 11(5):561–570.

Hernán, M. A., Brumback, B. A., and Robins, J. M. (2002). Estimating the causal effect of zidovudine on CD4 count with a marginal structural model for repeated measures. *Statistics in medicine*, 21(12):1689–1709.

Hernán, M. A. and Robins, J. M. (2006). Estimating causal effects from epidemiological data. *Journal of Epidemiology of Community Health*, 60(7):578–86.

Hernán, M. A. and Robins, J. M. (2017). *Causal Inference*. Chapman & Hall/CRC, forthcoming, Boca Raton.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, 81(396):945–960.

Howarth, R. W. and Marino, R. (2006). Nitrogen as the limiting nutrient for eutrophication in coastal marine ecosystems: evolving views over three decades. *Limnology and Oceanography*, 51(1part2):364–376.

Huber, P. J. (1964). Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35(1):73–101.

Huber, P. J. and Ronchetti, E. M. (2009). *Robust Statistics*. Wiley, Hoboken, 2nd edition.

Hudgens, M. G. and Halloran, M. E. (2006). Causal vaccine effects on binary postinfection outcomes. *Journal of the American Statistical Association*, 101(473):51–64.

Hudgens, M. G. and Halloran, M. E. (2008). Toward causal inference with interference. *Journal of the American Statistical Association*, 103(482):832–842.

Hudgens, M. G., Hoering, A., and Self, S. G. (2003). On the analysis of viral load endpoints in HIV vaccine trials. *Statistics in medicine*, 22(14):2281–2298.

Isaacs, J. D., Strangman, W. K., Barbera, A. E., Mallin, M. A., McIver, M. R., and Wright, J. L. (2014). Microcystins and two new micropeptin cyanopeptides produced by unprecedented *Microcystis aeruginosa* blooms in North Carolina's Cape Fear River. *Harmful Algae*, 31:82–86.

Jackson, L. A., Jackson, M. L., Nelson, J. C., Neuzil, K. M., and Weiss, N. S. (2006). Evidence of bias in estimates of influenza vaccine effectiveness in seniors. *International Journal of Epidemiology*, 35(2):337–344.

Jackson, M. L. and Nelson, J. C. (2013). The test-negative design for estimating influenza vaccine effectiveness. *Vaccine*, 31(17):2165–2168.

Joffe, M. (2011). Principal stratification and attribution prohibition: good ideas taken too far. *International Journal of Biostatistics*, 7(1):1–22.

Joffe, M., Gambhir, M., Chadeau-Hyam, M., and Vineis, P. (2012). Causal diagrams in systems epidemiology. *Emerging themes in epidemiology*, 9(1):1.

Kauermann, G. and Carroll, R. (2001). A note on the efficiency of sandwich covariance matrix estimation. *Journal of the American Statistical Association*, 96(456):1387–1396.

Kennedy, J. T. and Whalen, S. C. (2007). Seasonality and controls of phytoplankton productivity in the middle Cape Fear River, USA. *Hydrobiologia*, 598(1):203–217.

Li, P. and Redden, D. T. (2015). Small sample performance of bias-corrected sandwich estimators for cluster-randomized trials with binary outcomes. *Statistics in Medicine*, 34(2):281–96.

Liang, K.-Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1):13–22.

Lipsitch, M., Tchetgen Tchetgen, E. J., and Cohen, T. (2010). Negative controls: a tool for detecting confounding and bias in observational studies. *Epidemiology*, 21(3):383–388.

Lu, B., Preisser, J. S., Qaqish, B. F., Suchindran, C., Bangdiwala, S. I., and Wolfson, M. (2007). A comparison of two bias-corrected covariance estimators for generalized estimating equations. *Biometrics*, 63(3):935–941.

Mallin, M. A., Johnson, V. L., Ensign, S. H., and MacPherson, T. A. (2006). Factors contributing to hypoxia in rivers, lakes, and streams. *Limnology and Oceanography*, 51(1part2):690–701.

Mallin, M. A., McIver, M. R., Ensign, S. H., and Cahoon, L. B. (2004). Photosynthetic and heterotrophic impacts of nutrient loading to blackwater streams. *Ecological Applications*, 14(3):823–838.

Mancl, L. A. and DeRouen, T. A. (2001). A covariance estimator for GEE with improved small-sample properties. *Biometrics*, 57(1):126–134.

McDaniel, L. S., Henderson, N. C., and Rathouz, P. J. (2013). Fast pure R implementation of GEE: application of the Matrix package. *The R Journal*, 5:181–187.

Morgan, S. L. and Winship, C. (2014). *Counterfactuals and Causal Inference*. Cambridge University Press, 2nd edition.

NCDENR (2005). Cape Fear River basinwide water quality plan. Technical report, North Carolina Department of Environment and Natural Resources, Division of Water Quality/Planning, Raleigh, NC.

Norton, S. B., Cormier, S. M., and Suter, II, G. W. (2014). *Ecological Causal Assessment*. CRC Press.

Norton, S. B., Cormier, S. M., Suter, II, G. W., Schofield, K., Yuan, L., Shaw-Allen, P., and Ziegler, C. R. (2009). CADDIS: the causal analysis/diagnosis decision information system. In *Decision Support Systems for Risk-Based Management of Contaminated Sites*, pages 1–24.

Ogburn, E. L. and VanderWeele, T. J. (2014). Causal diagrams for interference. *Statistical Science*, 29(4):559–578.

Ogburn, E. L. and VanderWeele, T. J. (2017). Vaccines, contagion, and social networks. *Vaccines, Contagion, and Social Networks*, In press.

Paerl, H. W. (1987). Dynamics of blue-green algal (*Microcystis aeruginosa*) blooms in the lower Neuse River, NC: causative factors and potential controls. Technical Report 177.

Paul, S. and Zhang, X. (2014). Small sample GEE estimation of regression parameters for longitudinal data. *Statistics in Medicine*, 33(22):3869–81.

Pearl, J. (2009). Causal inference in statistics: An overview. *Statistics Surveys*, 3:96–146.

Pearl, J. (2010a). An introduction to causal inference. *International Journal of Biostatistics*, 6(2):1–62.

Pearl, J. (2010b). On the consistency rule in causal inference: Axiom, definition, assumption, or theorem? *Epidemiology*, 21(6):872–875.

Pearl, J. (2011). Principal stratification–a goal or a tool? *International Journal of Biostatistics*, 7(1):20.

Perez-Heydrich, C., Hudgens, M. G., Halloran, M. E., Clemens, J. D., Ali, M., and Emch, M. E. (2014). Assessing effects of cholera vaccination in the presence of interference. *Biometrics*, 70(3):731–741.

R Core Team (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rajbhandari, N. B., Pradhan, S., Di Luzio, M., Kebede, A., and Baker, V. (2015). Identifying nutrient contributors in North Carolina's coastal plain blackwater rivers. *American Journal of Environmental Sciences*, 11(4):313.

Richardson, D. B., Keil, A. P., Tchetgen Tchetgen, E., and Cooper, G. (2015). Negative control outcomes and the analysis of standardized mortality ratios. *Epidemiology*, 26(5):727–32.

Richardson, T. S. and Robins, J. M. (2013). Single world intervention graphs (SWIGs): A unification of the counterfactual and graphical approaches to causality. Technical Report 128, Center for Statistics and the Social Sciences, University of Washington.

Robins, J. (1986). A new approach to causal inference in mortality studies with a sustained exposure period application to control of the healthy worker survivor effect. *Mathematical Modelling*, 7(9-12):1393–1512.

Robins, J. M. (1989). The control of confounding by intermediate variables. *Statistics in medicine*, 8(6):679–701.

Robins, J. M. (1993). Analytic methods for estimating HIV-treatment and cofactor effects. In Kessler, R. C. and Ostrow, D. G., editors, *Methodological Issues in AIDS Behavioral Research*, pages 213–288. Plenum Press, New York.

Robins, J. M. (1997). Causal inference from complex longitudinal data. In Berkane, M., editor, *Latent Variable Modeling and Applications to Causality*, pages 69–117. Springer, New York.

Robins, J. M. and Greenland, S. (1992). Identifiability and exchangeability for direct and indirect effects. *Epidemiology*, 3:143–155.

Robins, J. M. and Hernán, M. A. (2009). Estimation of the causal effects of time-varying exposures. In Fitzmaurice, G., Davidian, M., Verbeke, G., and Molenberghs, G., editors, *Longitudinal Data Analysis*, pages 553–599. CRC Press, Boca Raton.

Robins, J. M., Hernán, M. A., and Brumback, B. (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology*, 11(5):550–560.

Robins, J. M. and Rotnitzky, A. (1992). Recovery of information and adjustment for dependent censoring using surrogate markers. In Jewell, N. P., Dietz, K., and Farewell, V. T., editors, *AIDS Epidemiology*, pages 297–331. Springer, New York.

Rubin, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *The Annals of Statistics*, 6(1):34–58.

Rubin, D. B. (1980). Randomization analysis of experimental data: The Fisher randomization test comment. *Journal of the American Statistical Association*, 75(371):591–593.

Rubin, D. B. (1986). Comment: Which ifs have causal answers. *Journal of the American Statistical Association*, 81(396):961–962.

Rubin, D. B. (2005). Causal inference using potential outcomes. *Journal of the American Statistical Association*, 100(469):322–331.

Saul, B. C., Hudgens, M. G., and Mallin, M. A. (2017). Upstream causes of downstream effects. *arXiv preprint arXiv:1705.07926*.

Schomaker, M., Egger, M., Ndirangu, J., Phiri, S., Moultrie, H., Technau, K., Cox, V., Giddy, J., Chimbetete, C., and Wood, R. (2013). When to start antiretroviral therapy in children aged 2–5 years: a collaborative causal modelling analysis of cohort studies from southern Africa. *PLoS Medicine*, 10(11):e1001555.

Schwartz, L. M., Halloran, M. E., Rowhani-Rahbar, A., Neuzil, K. M., and Victor, J. C. (2017). Rotavirus vaccine effectiveness in low-income settings: An evaluation of the test-negative design. *Vaccine*, 35(1):184–190.

Siegel, A., Cotti-Rausch, B., Greenfield, D. I., and Pinckney, J. L. (2011). Nutrient controls of planktonic cyanobacteria biomass in coastal stormwater detention ponds. *Marine Ecology Progress Series*, 434:15–27.

Sobel, M. E. (2006). What do randomized studies of housing mobility demonstrate?: Causal inference in the face of interference. *Journal of the American Statistical Association*, 101(476):1398–1407.

Soetaert, K. (2009). *rootSolve: Nonlinear root finding, equilibrium and steady-state analysis of ordinary differential equations*. R package version 1.6.

Soetaert, K. and Herman, P. M. (2009). *A practical guide to ecological modelling. Using R as a simulation platform*. Springer Science & Business Media.

Sofrygin, O., van der Laan, M. J., and Neugebauer, R. (2016). *simcausal: Simulating Longitudinal Data with Causal Inference Applications*. R package version 0.5.1.99.

Splawa-Neyman, J., Dabrowska, D. M., and Speed, T. P. (1990). On the application of probability theory to agricultural experiments. Essay on principles. Section 9. *Statistical Science*, 5(4):465–472. Translated from original appearing in *Annals of Agricultural Science* 1923.

Stefanski, L. A. and Boos, D. D. (2002). The calculus of M-estimation. *The American Statistician*, 56(1):29–38.

Sterne, J. A., Hernán, M. A., Ledergerber, B., Tilling, K., Weber, R., Sendi, P., Rickenbach, M., Robins, J. M., Egger, M., and Study, S. H. C. (2005). Long-term effectiveness of potent antiretroviral therapy in preventing AIDS and death: a prospective cohort study. *The Lancet*, 366(9483):378–384.

Sullivan, S. G., Tchetgen Tchetgen, E. J., and Cowling, B. J. (2016). Theoretical basis of the test-negative study design for assessment of influenza vaccine effectiveness. *American Journal of Epidemiology*, 184(5):345–353.

Suter, II, G. W., Norton, S. B., and Cormier, S. M. (2002). A methodology for inferring the causes of observed impairments in aquatic ecosystems. *Environmental Toxicology and Chemistry*, 21(6):1101–1111.

Tchetgen Tchetgen, E. J. and VanderWeele, T. J. (2012). On causal inference in the presence of interference. *Statistical Methods in Medical Research*, 21(1):55–75.

Teerenstra, S., Lu, B., Preisser, J. S., Van Achterberg, T., and Borm, G. F. (2010). Sample size considerations for GEE analyses of three-level cluster randomized trials. *Biometrics*, 66(4):1230–1237.

van der Laan, M. J. (2014). Causal inference for a population of causally connected units. *Journal of Causal Inference*, 2(1):13–74.

VanderWeele, T. J. (2009). Concerning the consistency assumption in causal inference. *Epidemiology*, 20(6):880–883.

VanderWeele, T. J. (2011). Principal stratification–Uses and limitations. *International Journal of Biostatistics*, 7(1):1–14.

VanderWeele, T. J., Tchetgen Tchetgen, E. J., and Halloran, M. E. (2012). Components of the indirect effect in vaccine trials: identification of contagion and infectiousness effects. *Epidemiology*, 23(5):751–61.

Vansteelandt, S. and Joffe, M. (2014). Structural nested models and G-estimation: The partially realized promise. *Statistical Science*, 29(4):707–731.

Verbitsky-Savitz, N. and Raudenbush, S. W. (2012). Causal inference under interference in spatial settings: A case study evaluating community policing program in Chicago. *Epidemiologic Methods*, 1(1):107–130.

Weisskopf, M. G., Tchetgen Tchetgen, E. J., and Raz, R. (2016). Commentary: On the use of imperfect negative control exposures in epidemiologic studies. *Epidemiology*, 27(3):365–7.

Westreich, D., Cole, S. R., Young, J. G., Palella, F., Tien, P. C., Kingsley, L., Gange, S. J., and Hernán, M. A. (2012). The parametric g-formula to estimate the effect of highly active antiretroviral therapy on incident AIDS or death. *Statistics in Medicine*, 31(18):2000–9.

Wickham, H. (2014). *Advanced R*. CRC Press.

Yuan, L. (2010). Estimating the effects of excess nutrients on stream invertebrates from observational data. *Ecological Applications*, 20(1):110–125.

Yuan, L., Pollard, A. I., Pather, S., Oliver, J. L., and D'Anglada, L. (2014). Managing microcystin: identifying national-scale thresholds for total nitrogen and chlorophyll *a*. *Freshwater Biology*, 59(9):1970–1981.

Zaadstra, B. M., Chorus, A. M. J., Van Buuren, S., Kalsbeek, H., and Van Noort, J. M. (2008). Selective association of multiple sclerosis with infectious mononucleosis. *Multiple Sclerosis*, 14(3):307–313.

Zeileis, A. (2004). Econometric computing with HC and HAC covariance matrix estimators. *Journal of Statistical Software*, 11(10):1–17.

Zeileis, A. (2006). Object-oriented computation of sandwich estimators. *Journal of Statistical Software*, 16(9):1–16.