Title of the dissertation

a thesis presented
by
Morgan F. Breitmeyer
to
The Departments of Statistics and Mathematics

in partial fulfillment of the requirements
for the degree of
Bachelor of the Arts
in the subject of
Statistics and Mathematics

Harvard University
Cambridge, Massachusetts
April 2017

©2017 – Morgan F. Breitmeyer
all rights reserved.

Thesis advisor: Professors James Robbins and Miguel Hernan Morgan F. Breitmeyer

Title of the dissertation

Abstract

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Morbi commodo, ipsum sed pharetra gravida, orci magna rhoncus neque, id pulvinar odio lorem non turpis. Nullam sit amet enim. Suspendisse id velit vitae ligula volutpat condimentum. Aliquam erat volutpat. Sed quis velit. Nulla facilisi. Nulla libero. Vivamus pharetra posuere sapien. Nam consectetuer. Sed aliquam, nunc eget euismod ullamcorper, lectus nunc ullamcorper orci, fermentum bibendum enim nibh eget ipsum. Donec porttitor ligula eu dolor. Maecenas vitae nulla consequat libero cursus venenatis. Nam magna enim, accumsan eu, blandit sed, blandit a, eros.

Quisque facilisis erat a dui. Nam malesuada ornare dolor. Cras gravida, diam sit amet rhoncus ornare, erat elit consectetuer erat, id egestas pede nibh eget odio. Proin tincidunt, velit vel porta elementum, magna diam molestie sapien, non aliquet massa pede eu diam. Aliquam iaculis. Fusce et ipsum et nulla tristique facilisis. Donec eget sem sit amet ligula viverra gravida. Etiam vehicula urna vel turpis. Suspendisse sagittis ante a urna. Morbi a est quis orci consequat rutrum. Nullam egestas feugiat felis. Integer adipiscing semper ligula. Nunc molestie, nisl sit amet cursus convallis, sapien lectus pretium metus, vitae pretium enim wisi id lectus. Donec vestibulum. Etiam vel nibh. Nulla facilisi. Mauris pharetra. Donec augue. Fusce ultrices, neque id dignissim ultrices, tellus mauris dictum elit, vel lacinia enim metus eu nunc.

Contents

1 Introduction 1

2 Background 5
- 2.1 Causal Effects . . . 6
- 2.2 Confounding . . . 9
- 2.3 Identifiability Assumptions . . . 10
- 2.4 IP Weighting . . . 13
- 2.5 Standardization . . . 15

3 Methods 17
- 3.1 Data Creation . . . 18
- 3.2 Parametric G-formula . . . 19
- 3.3 Doubly Robust Estimation . . . 23

4 Results Discussion 26

5 Conclusion 27

Appendix A Code 28

References 30

This is the dedication.

Acknowledgments

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Morbi commodo, ipsum sed pharetra gravida, orci magna rhoncus neque, id pulvinar odio lorem non turpis. Nullam sit amet enim. Suspendisse id velit vitae ligula volutpat condimentum. Aliquam erat volutpat. Sed quis velit. Nulla facilisi. Nulla libero. Vivamus pharetra posuere sapien. Nam consectetuer. Sed aliquam, nunc eget euismod ullamcorper, lectus nunc ullamcorper orci, fermentum bibendum enim nibh eget ipsum. Donec porttitor ligula eu dolor. Maecenas vitae nulla consequat libero cursus venenatis. Nam magna enim, accumsan eu, blandit sed, blandit a, eros.

1

Introduction

Causation versus correlation - the age old debate continues on among statisticians, scientists, and students alike. Correlation is the easier idea to understand here: are two (or more) things related to each other? Do taller people weigh more than shorter people? Is the temperature colder when there is snow on the ground than when there is not? Do people who drink red wine and eat dark

chocolate have healthier hearts?

Proving correlation is quite simple too. Find a group of people who do drink red wine and eat dark chocolate and a group of people who do not, and compare their resting heart rates and HDL ("good" cholesterol) levels. If the wine drinkers and chocolate consumers have better heart health, then it can be said that consuming these is associated with a healthier heart. As lovely as that sounds, there is a catch. This method only demonstrates that consuming red wine and chocolate is correlated with a healthy heart, but it does not prove that red wine and chocolate actually cause the healthy heart. The distinction being that perhaps everyone who consumes red wine and dark chocolate could also exercise frequently, be of a healthy weight, or be younger than those who do not, all possible factors that could lead to a healthier heart separate from an individual's red wine drinking and chocolate eating habits.

Yet, this begs the question of how to actually prove that the red wine and dark chocolate caused healthier hearts. Answering this question is not quite as simple as showing correlation, and traditionally, a fully randomized blinded experiment was required. A randomized trial experiment involves enrolling a sufficient number of patients, who are randomly categorized into one of two groups: those who are told to drink red wine (the treatment group) and those who are told not to (the controls). A trial like this would likely go on for some extended period of time, maybe a couple of months or even years. The individuals in both groups would have their heart rate and their HDL levels measured along the way. After many months of this, despite several individuals who surely had given up and dropped out of the study, data is acquired at the very end. Using this data, the researchers could compare the heart health of the two groups and come to a definitive conclusion

(pending statistical significance) of whether drinking red wine caused a healthier heart. This process is complicated, costly, and until recently, the only option to demonstrate causation.

This process is commonly used in medicine to test the efficacy of a treatment drug. On average, it takes 12 years and more than $350 million USD to get a new drug through clinical trial testing and FDA approval.[*] FDA approval requires the proof of statistically significant sufficient causation of the drug's efficacy through a randomized trial. The difficulty surrounding these trials makes the medical innovation process slow, expensive, and inefficient, purely because this means of proving causation is so tedious.

However, in the last few decades, a field of statistics known as causal inference has emerged, which studies and creates various methods to attempt to prove causation by better means than the randomized trial. This field emphasizes methods that can work strictly from observational data, meaning data that is already collected and does not need to be controlled or dictated by the investigator. Some examples of observational data include hospital data accumulated over millions of patients and across years, tracking symptoms, various levels, and outcomes, demographic surveys, and data from other experimental trials which is now being reconsidered for a different question of interest. Using these alternative methods, drug efficacy could be proven with significantly less cost and time, significantly impacting millions of lives.

This thesis is a study of two such methods of causal inference: the g-formula and a doubly robust estimator. Both methods seek to prove the same causal effect as a randomized trial from purely observational data. The doubly robust estimator is an improved development on the g-formula and is

[*] New Drug Approval Process, https://www.drugs.com/fda-approval-process.html

shown to be much more effective at correctly approximating causal effect. It can withstand significant error caused by human inaccuracy in model selection, leading to a better estimator.

2

Background

There's something to be said

2.1 Causal Effects

A traditional understanding of causation comes from the field of medicine, where researchers can perform a controlled experiment to prove causation. This type of study contains two sample groups, one which receives no treatment (the placebo group) and one which receives the treatment (the treatment group). Individuals are randomly allocated into one group, and by comparing the outcome of these two groups, the researchers can demonstrate whether the outcome for patients receiving treatment differs significantly from the controls. By quantifying the difference in outcomes between the groups, researchers can demonstrate an association between treatment and outcome. However, because of the randomized nature of the trial, association is causation.[*] [4]

To translate this idea into statistical terms, some notation must be introduced. The random variable A represents the treatment status, where a value of 1 indicates treated and a value of 0 indicates untreated. A fixed A, which has constant treatment over time, is written as A_i for the individual $i \in \{0, 1, 2, \ldots n\}$ with n the total number of individuals.[†] The random variable Y is the outcome variable, often with a value of 0 indicating survival and a value of 1 indicating death. These interpretations of A and Y correspond to the above understanding of causation studies, but for various causal inference studies, the form of Y and in particular can change depending on the question of interest. For example, Y can be a continuous variable, such as the weight difference of an individual in a weight loss trial or the change in HDL levels in a cholesterol study.

To study the causal effect of A, the desired value is the difference in Y under the varying condi-

[*] This idea is discussed further in Section 2.3.

[†] Non-fixed A representations are common and discussed in greater detail in Section 2.1.1.

tions of A. Notationally, this is the difference between $Y^{a=1}$ ‡, the outcome that would be observed under treatment, and $Y^{a=0}$, the outcome that would be observed under no treatment. This is in comparison to the observed outcome of Y or Y^A.

A causal effect can be seen on an individual level if $Y_i^{a=1} \neq Y_i^{a=0}$. By considering how each individual's responses to varying treatments differ, causation (or lack thereof) can easily be determined using paired differences of the form

$$Y_i^{a=1} - Y_i^{a=0} \tag{2.1}$$

These differences would be tested against the null hypothesis of zero difference in outcome for varying treatments.

However, certain difficulties arise using this method. In many studies, it is impossible to have scenarios of both treatment and no treatment for the same individual, particularly if a potential outcome is death. Typically, individuals either have $Y_i^{a=1}$ or $Y_i^{a=0}$, but not both, making it impossible to calculate the paired differences. Therefore, a controlled double blinded experiment is often performed, where each individual is randomly assigned treatment or placebo. In these studies, the statistic of interest is the average causal effect in the population,

$$\mathbb{E}[Y^{a=1}] - \mathbb{E}[Y^{a=0}] \tag{2.2}$$

‡Note that lowercase letters signify possible values of the random variable, in comparison to uppercase letters which represent actual observed values

Mathematically, this is equivalent to

$$\mathbb{E}[Y^{a=1} - Y^{a=0}] \tag{2.3}$$

because the average of differences is equal to the difference of averages.[5] Note, that this is not the same as calculating the mean of paired differences as if each individual had received both treatments at different times to calculate individual causal effects. Rather, the difference in the means of the placebo and treatment groups is being calculated to estimate average causal effect across the population.

2.1.1 Time Varying Data

Not all treatment regimens consist of constant treatment over a set period of time. Furthermore, if patients ina Therefore, a more complicated time-varying treatment can be considered. This would be written for a single individual as $\overline{A}_k = \{A_0, A_1, \cdots, A_k\}$, with time point $k \in \{0, 1, \ldots K\}$, given K as the maximum time value. The overline on $\overline{A}_k$ indicates the history of values up to and including time point k, and the notation $\overline{A}$ represents the full history. As an example of this, a patient with continuous treatment throughout the whole study would have data $\overline{A} = \{A_0 = 1, A_1 = 1, \ldots, A_K = 1\} = \{1, 1, \ldots 1\}$, which can also be written as $\overline{A} = \overline{1}$. In this scenario, the average causal effect is instead defined as

$$\mathbb{E}\left[Y^{\bar{a}=\bar{1}}\right] - \mathbb{E}\left[Y^{\bar{a}=\bar{0}}\right] \tag{2.4}$$

2.1.2 Deterministic Treatment Regimes

This time-varying framework of the treatment variable can also be considered for the covariate, L. From this, a dynamic treatment strategy can also be created, such that each possible realization $\bar{a}_k$ is dependent on the treatment and covariate history, $\bar{L}_k$ and $\bar{A}_{k-1}$. This can be written as a set of functions $\{g_k(\bar{a}_{k-1}, \bar{l}_k\}$ where g_k is the function making the treatment decision.

2.1.3 Sequentially Randomized Trial

A specific type of deterministic treatment regime is the sequentially randomized trial, in which a subject's treatment is chosen at each time from an associated density $f(a_k \mid \bar{l}_k, \bar{a}_{k-1})$ for $\bar{a}_k \in \overline{\mathcal{A}}_k$ where $\overline{\mathcal{A}}_k$ is the support of $\bar{A}_k$ in time period k.[12] In this type of randomized trial, each A_k for all subjects is chosen as an independent random draw from this type of distribution density. Sequentially randomized trials guarantee the identifiability assumptions of exchangeability and consistency to be discussed in Section 2.3.

2.2 Confounding

The use of the covariate $\bar{L}$ is a measurable proxy for an unmeasured and unknown underlying confounder, U. Theoretically, U should directly impact both L and Y, but not A, so it indicates a backdoor path between A and Y through U.[11] The expected way to account for the backdoor path caused by U would be to condition on it, but because it is unknown and therefore unmeasurable, this is not possible. Therefore, methods must be used to create this same effect using only L, which will allow

for the study of just the causal effect of A on Y. By eliminating the effect of U, there will no bias in the estimate of causal effect. The methods for doing so will be discussed in Sections 2.4 and 2.5.

In this scenario, L is referred to as a confounder for the effect of A, reflective of the fact that the underlying bias was the unknown of U and L is being used to account for that. It can be shown that in order to validly estimate the joint effect of all A_k simultaneously and without bias, it is sufficient (but not necessary) to block all backdoor paths from U to any A_k for all k.[8]

2.3 Identifiability Assumptions

It is sufficient to show that causal effects are valid and identifiable, meaning they have a single measurement of effect, on the following three assumptions: consistency, positivity, and exchangeability.[2,5] Under these three assumptions, the data closely resembles an ideal randomized trial. Through this, causation can be inferred, rather than simply association. Although the methods are directly testing association, these assumptions allow the tests to measure causation.

2.3.1 Consistency

Consistency is the idea that an individual's potential outcome and their observed outcome are equal[2,5]. Statistically, this is

$$\text{If } A_i = a, \text{ then, } Y_i^a = Y^{A_i} = Y_i \tag{2.5}$$

where Y_i^a is individual i's potential outcome and A_i is the observed treatment.

Consistency can deteriorate under the presence of multiple or varying treatment options, such as different surgeons perform a procedure or even varying procedures. Protection against this is partially in the understanding and reasonable pruning of the data. This can be done through clear and precise questions of interest, and hopefully, detailed data that allows for comprehensive refinement.

This idea can be expanded to time-varying treatment and covariate variables, as follows

$$\text{If } \overline{A}_k = \bar{a}_k^g, \text{ then, } \overline{Y}_{k+1} = \bar{Y}_{k+1}^g \text{ and } \overline{L}_k = \bar{L}_k^g \tag{2.6}$$

2.3.2 Exchangeability

Exchangeability is the idea that individuals in either group of a randomized experiment would have had the same response given the treatment.[5] There should be no bias to either group to respond favorably or not to treatment or lack thereof; thus, the results should be equivalent if any subject is moved from one group to the other.

Statistically, this is $P[Y^a = 1 \mid A = 1] = P[Y^a = 1 \mid A = 0] = P[Y^a = 1]$. This means that Y^a is independent of A, and the treatment has no predictive power of the outcome. This independence allows for several conclusions. Firstly, $E[Y^a \mid A = a'] = E[Y^a]$ by definition of independence.

Given some indicator of prognosis in the form of L, exchangeability is possible for those with similar prognoses, but it becomes problematic across varying prognoses. For example, exchangeability is attainable when considering obesity, but is more difficult when the confounder has a high mortality rate, such as Therefore, conditional exchangeability is obtained: $P[Y^a = 1 \mid A = a, L = l] = P[Y^a = 1 \mid A \neq a, L = l]$, i.e. $Y^a \perp\!\!\!\perp A \mid L$.[5] Conditional exchangeability guarantees the ability to measure

effects using complete data.

The power of a randomized trial is that it should theoretically create exchangeability. By randomly putting subjects into their groups, there should be no reason that the patients between the two groups differ or will respond to treatment differently. However, exchangeability can be obtained in an observational study if $P[A_k = 1]$ depends only on $\{\overline{A}_{k-1}, \overline{L}_k\}$ and thus,

$$P[A_k \mid \overline{A}_{k-1}, \overline{L}_k] \perp\!\!\!\perp U \tag{2.7}$$

By accounting for U using L in a time-varying treatment method, it can be seen that

$$Y \perp\!\!\!\perp A_k \mid \overline{L}_k, \overline{A}_{k-1} \tag{2.8}$$

which is referred to as having no unmeasured time-varying confounders. Although guaranteed for fixed treatments, sequential exchangeability is not guaranteed.[11] Approximate exchangeability can be achieved in practice by including as many covariates as is feasibly reasonable, but this is still risky business there is no known method for computationally measuring or empirically testing sequential exchangeability. However, the assumption of conditional exchangeability is the same as for fixed treatment models and is sufficient for determining causal effect.

2.3.3 Positivity

Positivity is the condition that a specified conditional probability is well-defined, meaning that for every value of the covariate L, there exist subjects with a specified value of a.[4] Statistically, this looks like

$$P[A = a \mid L = l] > 0 \quad \forall l, \text{ such that } P[L = l] \neq 0 \tag{2.9}$$

This can also be expressed for time-varying treatments as follows,

$$P[A_k = a_k \mid \overline{L}_k, \overline{A}_{k-1}] > 0 \quad \forall A_k, \text{ such that } P[\overline{L}_k = \overline{l}_k, \overline{A}_{k-1} = \overline{a}_{k-1}] \neq 0 \tag{2.10}$$

2.4 IP Weighting

Many of the concerns discussed above can by addressed using the method of IP weighting by simulating a pseudo-population, in which every individual has two data inputs, the expected observed outcomes under treatment and under no treatment. The method by which this is done is by considering a confounder of the data, L, a value which is known before treatment and often factors into the decision to assign treatment. For example, a confounder in a study on a cholesterol drug could be whether the patient is obese or has high blood pressure. By creating the pseudo-population, the treatment and placebo groups share the same underlying covariate characterizations and distributions.

The pseudo-population can be calculated with the following for each of the possible A and L combinations

$$n \cdot P[Y = y \mid A = a, L = 1] \cdot P[A = a \mid L = l] \cdot P[L = l] \cdot \frac{1}{P[A = a \mid L = 1]} \tag{2.11}$$

where the last term here is the IP weight, $W^A = 1/f(A|L)$.

This weight is equivalent to the inverse of the propensity score, which can be defined as the probability of receiving treatment and written as,[6]

$$e(x) = \frac{N_t(x)}{N_c(x) + N_t(x)} = P[A = a \mid L = l] \tag{2.12}$$

where $x = X_i$ is the population data and $N_t(x)$ and $N_c(x)$ are the number of individuals in the treatment and control groups respectively.

This form in expression 2.11 can be used to solve for the standardized mean as follows,

$$E[Y^a] = \sum_l n \cdot P[Y = y \mid A = a, L = 1] \cdot P[A = a \mid L = l] \cdot P[L = l] \cdot \frac{1}{P[A = a \mid L = 1]} \tag{2.13}$$

$$= \sum_l n \cdot P[Y = y \mid A = a, L = 1] \cdot P[L = l] \tag{2.14}$$

$$= \sum_l E[Y \mid A = a, L = l] P[L = l] \tag{2.15}$$

This leads to the confounders being accounted for or eliminated in the pseudo-population. As a result, the causal effect of A on Y can effectively be estimated using the pseudo-population without

any impact from the confounders.

2.4.1 Parametric Estimates

The above non-parametric values for $P[A = a \mid L = l]$ are effective for limited dichotomous confounders, but this method has limitations when L is highly dimensional. To address this, a parametric estimate $\widehat{P}[A = a \mid L = l]$ can be obtained using a logistic regression model for A with all the confounders in L included as covariates. This allows us to estimate IP weights

2.5 Standardization

Like IP weighting, standardization is a method of calculating the marginal counterfactual risk of $P[Y^a = 1]$. This method weights the population by conditioning on the covariates levels in L, in order to make the probability of treatment A independent of the covariates. The weighting looks like this,

$$P[Y^a = 1] = \sum_l P[Y^a = 1 \mid L = l] \cdot P[L = l] \tag{2.16}$$

$$= \sum_l P[Y = 1 \mid L = l] P[L = l] \tag{2.17}$$

where the equality is because of the conditional exchangeability. This standardization method can be used to obtain the standardized mean,

$$E[Y^a] = E[Y \mid L = l, A = a] \cdot P[L = l] \tag{2.18}$$

Note that this returns the same non-parametric expression for the standardized mean as the method of IP weighting because they are mathematically equivalent.

3

Methods

Two formal methods for estimating, causal effect were used in this study: g-formula estimation and doubly robust estimation. These two methods are developments of standardization and IP weighting as discussed in the previous section. They were implemented and studied using simulated data according to the following methods.

3.1 Data Creation

For the purposes of this study, a data generating algorithm was created to provide consistent and easily accessible data for many simulations. The data generated was time-varying and sequentially randomized according to the following schematic.

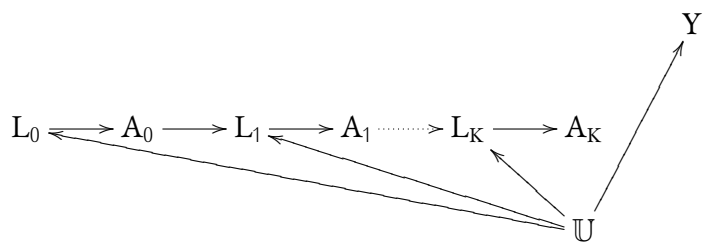

Figure 3.1: Diagram of conditional dependencies in data generating process.

The algorithm to generate datasets is as follows for each individual, of which 1,000 were simulated in this study.

1. Take the predetermined coefficients $\vec{\alpha}$ and $\vec{\beta}$, which are generated outside the data generating process to provide consistency. In this study, the values were as follows,

$\vec{\alpha}$	$\vec{\beta}$
$\alpha_0 = 0.58986656$	$\beta_0 = 0.17868818$
$\alpha_1 = 0.95344212$	$\beta_1 = 0.89069712$
$\alpha_2 = -0.89822429$	$\beta_2 = 0.89037635$
$\alpha_3 = -0.95566697$	$\beta_3 = 0.20497534$
$\alpha_4 = 0.67520365$	$\beta_4 = 0.10442911$
$\alpha_5 = 2.46365403$	

   These coefficients were created by pulling each from $\vec{\alpha} \sim \text{Uniform}(-1.0, 1.0)$ and $\vec{\beta} \sim \text{Uniform}(-1.0, 1.0)$. The one change made was that α_5 had 1.5 added to the randomly generated value to ensure that the underlying covariate had significant impact.

2. Create the underlying confounder, U_i from $U_i \sim \text{Unif}(0.1, 1)$

3. Using the following logistic expressions, probabilities for each $L_{k,i}$ and $A_{k,i}$ can be obtained where k is the time and i is the individual, conditional on $\overline{L}_{k-1,i}$ and $\overline{A}_{k-1,i}$. These probabilities are then used to obtain values for $L_{k,i}$ and $A_{k,i}$ using a binomial distribution with the respective probabilities.

$$\text{logit}P[L_{k,i}] = \alpha_0 + \alpha_1 \cdot L_{k-1,i} + \alpha_2 \cdot L_{k-2,i} + \alpha_3 A_{k-1,i} + \alpha_4 A_{k-2,i} + \alpha_5 U_i \tag{3.1}$$

$$\text{logit}[A_{k,i}] = \beta_0 + \beta_1 L_{k,i} + \beta_2 L_{k-1,i} + \beta_3 A_{k-1,i} + \beta_4 A_{k-2,i} \tag{3.2}$$

Note that for low time values when history is limited, the above expressions are slightly modified as follows,

$$\text{logit}P[L_{0,i}] = \alpha_0 + \alpha_5 U_i \tag{3.3}$$

$$\text{logit}[A_{0,i}] = \beta_0 + \beta_1 L_{k,i} \tag{3.4}$$

$$\text{logit}P[L_{1,i}] = \alpha_0 + \alpha_1 \cdot L_{k-1,i} + \alpha_3 A_{k-1,i} + \alpha_5 U_i \tag{3.5}$$

$$\text{logit}[A_{1,i}] = \beta_0 + \beta_1 L_{k,i} + \beta_2 L_{k-1,i} + \beta_3 A_{k-1,i} \tag{3.6}$$

4. Obtain a final Y_i value for each individual where $Y_i \sim \mathcal{N}(\mu = U, \sigma = 1)$

3.2 Parametric G-formula

Similar to IP weighting, parametric estimates can be obtained for standardized estimates. An efficient method for doing this is the generalization of standardization to time-varying treatments and confounders, coined the g-formula method by Robins in 1986.[9,11,5] The method can be used for fixed and time-varying treatments in longitudinal studies, and it seeks to estimate the average causal effect of treatment.

which can be estimated as

$$\mathbb{E}[Y^{\bar{a}=\bar{1}}] - \mathbb{E}[Y^{\bar{a}=\bar{0}}] \tag{3.7}$$

where the respective $\bar{a} = \bar{1}$ and $\bar{a} = \bar{0}$ signify constant treatment and no treatment over the entire time period.

The g-formula seeks to calculate each standardized mean using the following,

$$\mathbb{E}[Y^{\bar{a}=\bar{1}}] = \sum_{l_i} \mathbb{E}[Y \mid \bar{L}_t, \bar{A}_t] \cdot \prod_{k=0}^{t} P[L_k = l_k \mid \bar{L}_{k-1}, \bar{A}_{k-1}] \tag{3.8}$$

where $\bar{L}_k = \{L_0 = l_0, L_1 = l_1, \cdots, L_k = l_k\}$ and $\bar{A}_k = \{a_0 = 1, a_1 = 1, \cdots, a_k = 1\}$ are the history of the treatment and covariate variables up to and including time k. The equivalent formula can be derived for $\mathbb{E}[Y^{\bar{a}=\bar{0}}]$. In expression 3.10, the summation term is

One of the key reasons for using the g-formula method is that it is able to account for time-varying confounders which have feedback to each other. This is equivalent to each L_k being dependent on A_{k-1}.[9] In these scenarios, traditional methods for adjusting for the confounder, such as stratification, regression, and matching, may introduce bias; however, the g-formula method (as well as IP weighting) will not.[11] This is because these other methods are unable to estimate the joint effect of all treatment values $\{A_0, A_1 \ldots A_K\}$ simultaneously and without bias.[3]

The g-formula method has been shown to have a smaller variance than IP weighting methods, but this comes with added parametric modeling assumptions.[12] The smaller variance is due to the

fact that the g-formula uses maximum likelihood estimates, in comparioson to the semi-parametic estimator used in IP weighting. Furthermore, IP weighting does fault and become quite unstable under violations (or close violations) of the positivity assumption, due to division by a near zero probability $P[A_k = a_k \mid \overline{L}_k, \overline{A}_{k-1}]$.

These improvements are, however, under the assumption of exchangeability, and the fact that the g-formula relies more heavily on parametric assumptions, which can lead to bias. The presence of bias is dependent on the accuracy of the models for Y, A_k and L_k for all k. IP weighting methods are also dependent on the accuracy of their models, just different models such as for A_k conditional on $\overline{L}_k, \overline{A}_{k-1}$.

3.2.1 Protocol

The method is performed in several steps, as follows

1. <u>Create outcome models</u>: Create models for the outcome variable Y and the covariates, L_k at each time using the original dataset. The model for Y is regressed on the treatment variable A and the confounders, L.

   In this case, the following models were chosen for $Y \mid \overline{A}_t, \overline{L}_t$ and $L_k \mid \overline{L}_{k-1}, \overline{A}_{k-1}$,

$$\mathbb{E}[Y \mid \overline{A}_t, \overline{L}_t] = \theta_0 + \theta_1 A_t + \cdots + \theta_j A_0 + \theta_{j+1} L_t + \cdots + \theta_{j+k} L_0 \tag{3.9}$$

$$\text{logit}[L_k \mid \overline{L}_{k-1}, \overline{A}_{k-1}] = \gamma_0 + \gamma_1 L_{k-1} + \gamma_2 L_{k-2} + \gamma_3 L_{k-3} + \gamma_4 A_{k-1} + \gamma_5 A_{k-2} + \gamma_6 A_{k-3} \tag{3.10}$$

   A time lag of only three historical values was deemed sufficient for the model of L_k because ...

   Note that for initial time points where there was insufficient history for the full model,

smaller models were created as follows

$$\text{Logit}[L_1 \mid L_0, A_0] = \gamma_0' + \gamma_1' L_0 + \gamma_2' A_0 \tag{3.11}$$

$$\text{Logit}[L_2 \mid L_0, L_1, A_0, A_1] = \gamma_0'' + \gamma_1'' L_{k-1} + \gamma_2'' L_{k-2} + \gamma_3'' A_{k-1} + \gamma_4'' A_{k-2} \tag{3.12}$$

2. Predict using Monte Carlo: Using the model created in step 2, predict the outcome Y for the two new datasets created in step 1, conditioned on the given A and L values.

   Using expressions 3.9 through 3.12, a Monte Carlo simulation must be performed to gain prediction values. This is because it is impractical to calculate expression 3.10 directly for a continuous L. This process is done as follows for time $k = \{0, \ldots, K\}$, and $i = \{1, \ldots, n\}$ keeping the test treatment regimen of interest $\bar{a}$ in mind through the process.

   (a) Select the L_0 value from a random individual from $i \in \{1, \ldots, n\}$.

   (b) Obtain a probability of L_1 using this L_0 and a_0 in expression 3.11 and then obtain a sample L_1 by pulling from a binomial distribution.

   (c) Obtain a probability of L_2 using the L_0, L_1, a_0, and a_1 in expression 3.12 and then obtain a sample L_2 by pulling from a binomial distribution.

   (d) Continue the above process until time K using expression 3.10 to get a full history $\bar{L}_K$ and all the probabilities $P[L_k = l_k \mid \bar{L}_{k-1}, \bar{A}_{k-1}]$

   (e) Using expression 3.9, $\bar{a}$ and the above solved for $\bar{L}_K$, calculate $\mathbb{E}[Y \mid \bar{A}_t, \bar{L}_t]$

   (f) Take the product of all the probabilities $P[L_k = l_k \mid \bar{L}_{k-1}, \bar{A}_{k-1}]$ for $k = 0, \ldots, K$ and $\mathbb{E}[Y \mid \bar{A}_t, \bar{L}_t]$ to get a final estimate.

   (g) Repeat steps (2a) through (2f) for as many simulations as desired. In this study, 10,000 individuals were simulated.

   (h) Take the mean of all simulation values to obtain $\mathbb{E}[Y^{\bar{a}}]$

   (i) Repeat all above steps for the opposing treatment regimen of interest $\bar{a}'$ and take difference $\mathbb{E}[Y^{\bar{a}}] - \mathbb{E}[Y^{\bar{a}'}]$ to get the average causal treatment effect.

To do this process using non-parametric estimates, the Monte Carlo simulation is unnecessary. Instead, the methodology would be to create two new simulated datasets, the first having all individuals under no treatment ($A = 0$) and the second having all individuals treated ($A = 1$). Each of

these new datasets has the same size as the original and the same "individuals", just changed values for A. For these two datasets, delete the outcome values for Y to treat it as a missing data. Then, the outcome models would be calculated as above. The prediction step, however, would differ, instead using the models directly to get predicted values for each individual in the extra two datasets. Finally, the standardized means could be obtained by creating a weighted average for $E[Y^{a=0}]$ from the first new dataset and one for $E[Y^{a=1}]$ from the second new dataset.

Variance Estimate

3.3 Doubly Robust Estimation

The method of doubly robust estimation, as proposed by Bang and Robins[1], combines the two previously discussed methods of IP weighting and standardization.

IP weighting and standardization techniques are expected to provide different answers, unless there are no models used to create estimates.[5] IP weighting estimates $P[A = a \mid L = l]$ using $P[A = a \mid L = l]$, while standardization estimates $E[Y \mid A = a, L = l]$

- does not use the observed data treatment density

- estimators are consistent if either the model for treatment given the past (as in IP weighting) is correctly specified or the models for the outcome and covariates given the past (as needed to implement the parametric g-formula) are correctly specified, without knowing which is correct

-

3.3.1 Protocol

The method can be performed recursively using the following steps,

1. Build a model for the treatment A_k with data pooled for all time $m \in \{1, \ldots, K\}$ and all individuals $i \in \{1, \ldots n\}$ and obtain the MLE $\hat{\alpha}$ of α using logistic regression.

$$\text{logit}\{P[A_{m,i} = 1 \mid \bar{l}_{m,i}, \bar{a}_{m-1,i}; \alpha]\} = w_m[\bar{l}_{m,i}, \bar{a}_{m-1,i}; \alpha] \tag{3.13}$$

   This model can be rewritten as the following.

$$f(A_m \mid \overline{L}_m, \overline{A}_{m-1}; \hat{\alpha}) = \alpha_0 + \alpha_1 \cdot L_m + \alpha_2 \cdot A_{m-1} + \alpha_3 \cdot L_{m-1} + \alpha_4 \cdot L_{m-2} + \alpha_5 \cdot A_{m-2} \tag{3.14}$$

2. Set $\hat{T}_{K+1} = Y$

3. Recurse for $m = K + 1, \ldots, 2$

   (a) Use IRLS and a specified parametric regression model to get

$$h_{m-1}(\overline{L}_{m-1}, \overline{A}_{m-1}; \beta_{m-1}, \phi_{m-1}) = \Psi\{s_{m-1}(\overline{L}_{m-1}, \overline{A}_{m-1}; \beta_{m-1}) + \phi_{m-1}\overline{\pi}_{m-1}^{-1}(\hat{\alpha})\} \tag{3.15}$$

   which gives the conditional expectation of

$$\mathbb{E}\left[\hat{T}_m \mid \overline{L}_{m-1}, \overline{A}_{m-1}\right] \tag{3.16}$$

   The known function s_m is specified on a case by case basis, and in this case was chosen to be as follows for the unknown parameter β.

$$s_m(\overline{L}_m, \overline{A}_m; \beta_m) = \theta_0 + \theta_1 L_m + \theta_2 A_m + \theta_3 L_{m-1} + \theta_4 A_{m-1} + \theta_5 L_{m-2} + \theta_6 A_{m-2} \tag{3.17}$$

   Furthermore, the function $\overline{\pi}_m(\hat{\alpha})$ is the propensity score model and is specified as follows

$$\overline{\pi}_m(\hat{\alpha}) = \prod_{j=1}^{m} f(A_m \mid \overline{L}_m, \overline{A}_{m-1}; \hat{\alpha}) \tag{3.18}$$

$$= \zeta_0 + \zeta_1 L_m + \zeta_2 L_{m-1} + \zeta_3 A_{m-1} + \zeta_4 A_{m-2} \tag{3.19}$$

The given Ψ is the canonical link function of the chosen GLM. *

(b) Let $\hat{h}_{m-1}(\overline{L}_{m-1}, \overline{A}_{m-1}; \hat{\beta}_{m-1}, \hat{\phi}_{m-1})$ be the predicted model derived in step 3a. This implies that $(\hat{\beta}'_{m-1}, \hat{\phi}'_{m-1})$ is a solution of

$$0 = \tilde{\mathbb{E}}\left[\left[\hat{T}_m - \Psi\{s_{m-1}(\overline{L}_{m-1}, \overline{A}_{m-1}; \hat{\beta}_{m-1}) + \hat{\phi}_{m-1}\overline{\pi}^{-1}_{m-1}(\hat{\alpha})\}\right]\left(\frac{\partial s(\overline{L}_{m-1}; \beta_{m-1})}{\partial \beta'_{m-1}, \overline{\pi}^{-1}_{m-1}(\hat{\alpha})}\right)\right] \tag{3.20}$$

where $\tilde{\mathbb{E}}(X) = \frac{1}{n}\sum_{i=1}^{n} X_i$ is the computational average.

(c) Set

$$\hat{T}^{a_{m-1},\ldots,a_K}_{m-1} = \hat{h}_{m-1}(\overline{L}_{m-1}, \overline{A}_{m-2}, a_{m-1}) \tag{3.21}$$

$$= \Psi\{s_{m-1}(\overline{L}_{m-1}, \overline{A}_{m-2}, a_{m-1}; \beta_{m-1}) + \phi_{m-1}\overline{\pi}^{-1}_{m-2}(\hat{\alpha})f(a_{m-1} \mid \overline{L}_{m-1}, \overline{A}_{m-2}; \hat{\alpha})\} \tag{3.22}$$

where a_{m-1} is our treatment value of interest, the lowercase letter indicating a test value rather than an observed.

4. To calculate the final $\mathbb{E}[Y^{\bar{a}}]$, solve

$$\mathbb{E}[Y^{\bar{a}}] = \tilde{\mathbb{E}}(\hat{T}_1) = \tilde{\mathbb{E}}(\hat{T}^{\bar{a}}_1) \tag{3.23}$$

5. Repeat all above steps for the opposing treatment regimen of interest $\bar{a}'$ and take difference $\mathbb{E}[Y^{\bar{a}}] - \mathbb{E}[Y^{\bar{a}'}]$ to get the average causal treatment effect.

*The desired method to do this is using a GLM with an underlying distribution (or family) of a Gaussian normal and a logit link. However, python does not have the capacity to do it in this way, so alternatives had to be tested and considered, including basic linear regression with an expit applied after step ?? as well as using a logistic regression and taking the predicted probability to pull 1000 samples for each individual and regressing off that new data in the next step.

4

Results Discussion

Lorem ipsum dolor sit amet,

5

Conclusion

A

Code

References

[1] Bang, H. & Robins, J. M. (2005). Doubly robust estimation in missing data and causal inference models. Biometrics, 61(4), 962–973.

[2] Cole, S. R. & Frangakis, C. E. (2009). The consistency statement in causal inference: a definition or an assumption? Epidemiology, 20(1), 3–5.

[3] Fitzmaurice, G., Davidian, M., Verbeke, G., & Molenberghs, G. (2008). Longitudinal data analysis. CRC Press.

[4] Hernán, M. A. & Robins, J. M. (2006). Estimating causal effects from epidemiological data. Journal of epidemiology and community health, 60(7), 578–586.

[5] Hernan, M. A. & Robins, J. M. (2016). Causal Inference. Chapman & Hall/CRC.

[6] Imbens, G. W. & Rubin, D. B. (2015). Causal inference in statistics, social, and biomedical sciences. Cambridge University Press.

[7] Lodi, S., Phillips, A., Logan, R., Olson, A., Costagliola, D., Abgrall, S., van Sighem, A., Reiss, P., Miró, J. M., Ferrer, E., et al. (2015). Comparative effectiveness of immediate antiretroviral therapy versus cd4-based initiation in hiv-positive individuals in high-income countries: observational cohort study. The Lancet HIV, 2(8), e335–e343.

[8] Pearl, J. & Robins, J. (1995). Probabilistic evaluation of sequential plans from causal models with hidden variables. In Proceedings of the Eleventh conference on Uncertainty in artificial intelligence (pp. 444–453).: Morgan Kaufmann Publishers Inc.

[9] 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.

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

[11] Wright, J. D. (2015). International encyclopedia of the social and behavioral sciences.

[12] Young, J. G., Cain, L. E., Robins, J. M., O'Reilly, E. J., & Hernán, M. A. (2011). Comparative effectiveness of dynamic treatment regimes: an application of the parametric g-formula. Statistics in biosciences, 3(1), 119–143.