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

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

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.

0

Introduction

1

Background

There's something to be said for having a good opening line. Morbi commodo, ipsum sed pharetra gravida, orci $x = 1/\alpha$ magna rhoncus neque, id pulvinar odio lorem non turpis[2,4].

1.1 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.[*] ?

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. 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 conditions of A. Notationally, this is the difference between $Y^{a=1}$, the outcome that would be observed

[*] This idea is discussed further in Section 1.2.

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 $Y_i^{a=1} \neq Y_i^{a=0}$ for individual i. 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

$$Y_i^{a=1} - Y_i^{a=0} \tag{1.1}$$

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,

$$\mathbb{E}[Y^{a=1}] - \mathbb{E}[Y^{a=0}] \tag{1.2}$$

Mathematically, this is equivalent to

$$\mathbb{E}[Y^{a=1} - Y^{a=0}] \tag{1.3}$$

because the average of differences is equal to the difference of averages.[3] 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.

1.2 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.[?] [3] 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.

1.2.1 Consistency

Consistency is the idea that an individual's potential outcome and their observed outcome are equal[?] [3]. Statistically, this is

$$\text{if } A_i = a, \text{ then, } Y_i^a = Y^{A_i} = Y_i \tag{1.4}$$

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.

1.2.2 Exchangeability

Exchangeability is the idea that individuals in either group of a randomized experiment would have had the same response given the treatment. [3]There should be no bias to either group to respond favorably or not to treatment or lack thereof. 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 progonosis in the form of L, exchangeability is possible for those with

similar prognoses, but problematic across varying prognoses. 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. $Y^a \perp\!\!\!\perp A \mid L$.[3] 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.

1.2.3 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.[?] Statistically, this looks like

$$P[A = a \mid L = l] > 0 \tag{1.5}$$

for a l with $P[L = l] \neq 0$.

1.3 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

$$n \cdot P[Y = y \mid A = a, L = l] \cdot P[A = a \mid L = l] \cdot P[L = l] \cdot \frac{1}{P[A = a \mid L = l]} \tag{1.6}$$

where the last term here is the IP weight, $W^A = 1/f(A|L)$. This form can be used to solve for the standardized mean as follows,

$$E[Y^a] = \sum_l n \cdot P[Y = y \mid A = a, L = l] \cdot P[A = a \mid L = l] \cdot P[L = l] \cdot \frac{1}{P[A = a \mid L = l]} \tag{1.7}$$

$$= \sum_l n \cdot P[Y = y \mid A = a, L = l] \cdot P[L = l] \tag{1.8}$$

$$= \sum_l E[Y \mid A = a, L = l] P[L = l] \tag{1.9}$$

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.

1.3.1 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

1.4 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.

$$P[Y^a = 1] = \sum_l P[Y^a = 1 \mid L = l] \cdot P[L = l] \tag{1.10}$$

$$= \sum_l P[Y = 1 \mid L = l] P[L = l] \tag{1.11}$$

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

$$E[Y^a] = E[Y \mid L = l, A = a] \cdot P[L = l] \tag{1.12}$$

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

1.5 Parametric G-formula

1.6 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.

2

Methods

Lorem ipsum dolor sit amet,

3

Simulation Discussion

Lorem ipsum dolor sit amet,

4

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] Eigen, M. (1971). Selforganization of matter and the evolution of biological macromolecules. Naturwissenschaften, 58(10), 465–523.

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

[4] Knuth, D. E. (1968). Semantics of context-free languages. Mathematical Systems Theory, 2(2), 127–145.

[5] 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.

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

[7] 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.