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

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.[*3]

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 1.3.

[†] Non-fixed A representations are common and discussed in greater detail in Section 1.1.1.

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

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

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,

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

‡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

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

because the average of differences is equal to the difference of averages.[4] 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.

1.1.1 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 $\overline{A} = \bar{1}$. In this scenario, the average causal effect is instead defined as

$$\mathbb{E}\left[Y^{\bar{a}=\bar{1}}\right] - \mathbb{E}\left[Y^{\bar{a}=\bar{0}}\right] \tag{1.4}$$

1.1.2 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, $\overline{L}_k$ and $\overline{A}_{k-1}$. This can be written as a set of functions $\{g_k(\bar{a}_{k-1}, \overline{l}_k\}$ where g_k is the function making the treatment decision.

1.1.3 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 $f(a_k \mid \overline{l}_k, \bar{a}_{k-1})$ for $\bar{a}_k \in \overline{\mathcal{A}}_k$ where $\overline{\mathcal{A}}_k$ is the support of $\overline{A}_k$ in time period k.[10] 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 1.3.

1.2 Confounding

The use of the covariate $\overline{L}$ 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.[9] 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 1.4 and 1.5.

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.[?]

1.3 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.[2,4] 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.

1.3.1 Consistency

Consistency is the idea that an individual's potential outcome and their observed outcome are equal[2,4]. Statistically, this is

$$\text{If } A_i = a, \text{ then, } Y_i^a = Y^{A_i} = Y_i \tag{1.5}$$

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

$$\text{If } \overline{A}_k = \bar{a}_k^g, \text{ then, } \overline{Y}_{k+1} = \bar{Y}_{k+1}^g \text{ and } \overline{L}_k = \bar{L}_k^g \tag{1.6}$$

1.3.2 Exchangeability

Exchangeability is the idea that individuals in either group of a randomized experiment would have had the same response given the treatment.[4] 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. $Y^a \perp\!\!\!\perp A \mid L$.[4] 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,

$$P[A_k \mid \overline{A}_{k-1}, \overline{L}_k] \perp\!\!\!\perp U \tag{1.7}$$

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

$$Y \perp\!\!\!\perp A_k \mid \overline{L}_k, \overline{A}_{k-1} \tag{1.8}$$

which is referred to as having no unmeasured time-varying confounders. Although guaranteed for fixed treatments, sequential exchangeability is not guaranteed.[9] 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.

1.3.3 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.[3] Statistically, this looks like

$$P[A = a \mid L = l] > 0 \quad \forall l, \text{ such that } P[L = l] \neq 0 \tag{1.9}$$

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

$$P[A_k = a_k \mid \overline{L}_k, \overline{A}_{k-1}] > 0 \quad \forall A_k, \text{ such that } P[\overline{L}_k = \overline{l}_k, \overline{A}_{k-1} = \overline{a}_{k-1}] \neq 0 \tag{1.10}$$

1.4 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

$$n \cdot P[Y = y \mid A = a, L = l] \cdot P[A = a \mid L = l] \cdot P[L = l] \cdot \frac{1}{P[A = a \mid L = l]} \tag{1.11}$$

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,[5]

$$e(x) = \frac{N_t(x)}{N_c(x) + N_t(x)} = P[A = a \mid L = l] \tag{1.12}$$

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 equation 1.11 can be used to solve for the standardized mean as follows,

$$E[Y^a] = \sum_l n \cdot P[Y = y \mid A = a, L = l] \cdot P[A = a \mid L = l] \cdot P[L = l] \cdot \frac{1}{P[A = a \mid L = l]} \tag{1.13}$$

$$= \sum_l n \cdot P[Y = y \mid A = a, L = l] \cdot P[L = l] \tag{1.14}$$

$$= \sum_l E[Y \mid A = a, L = l] P[L = l] \tag{1.15}$$

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.

1.4.1 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

1.5 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,

$$P[Y^a = 1] = \sum_l P[Y^a = 1 \mid L = l] \cdot P[L = l] \tag{1.16}$$

$$= \sum_l P[Y = 1 \mid L = l] P[L = l] \tag{1.17}$$

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

$$E[Y^a] = E[Y \mid L = l, A = a] \cdot P[L = l] \tag{1.18}$$

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

2

Methods

Lorem ipsum dolor sit amet,

2.1 Data Creation

2.2 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.[7,9,4] 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

$$\mathbb{E}[Y^{\bar{a}=\bar{1}}] - \mathbb{E}[Y^{\bar{a}=\bar{0}}] \tag{2.1}$$

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,

$$\mathbb{E}[Y^{\bar{a}=\bar{1}}] = \sum_{l_i} \mathbb{E}[Y \mid \bar{L}_t, \bar{A}_t] \cdot \prod_{k=0}^{t} P[L_k = l_k \mid \bar{L}_{k-1}, \bar{A}_{k-1}] \tag{2.2}$$

where $\bar{L}_k = \{L_0 = l_0, L_1 = l_1, \cdots, L_k = l_k\}$ and $\bar{A}_k = \{a_0 = 1, a_1 = 1, \cdots, a_k = 1\}$. The equivalent formula can be derived for $\mathbb{E}[Y^{\bar{a}=\bar{0}}]$. In equation 2.2, 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}.[7] 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.[9] 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.[?]

The g-formula method has been shown to have a smaller variance than IP weighting methods, but this comes with added parametric modeling assumptions.[10]

2.2.1 Protocol

The method is performed in several steps, as follows

1. Expand the dataset: Create two new simulated datasets, the first has all individuals under no treatment ($A = 0$) and the second of which has 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.

2. Create outcome models: Create models for the outcome variable Y and the covariates, L_k at each time using the original dataset and the two datasets created in step 1 together. The model for Y is regressed on the treatment variable A and the confounders, L. Note that because the data in these two new datasets is missing for Y, they will not actually contribute to the model's parameters.

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

$$\mathbb{E}[Y \mid \overline{A}_t, \overline{L}_t] = \theta_0 + \theta_1 A_t + \cdots + \theta_j A_0 + \theta_{j+1} L_t + \cdots + \theta_{j+k} L_0 \tag{2.3}$$

$$\text{logit}[L_k \mid \overline{L}_{k-1}, \overline{A}_{k-1}] = \gamma_0 + \gamma_1 L_{k-1} + \gamma_2 L_{k-2} + \gamma_3 A_{k-1} + \gamma_4 A_{k-2} \tag{2.4}$$

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

3. Predict: 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 equation 2.3 above,

4. Standardization by averaging: Created 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

2.3 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.[4] IP weighting estimates $P[A = a \mid L = l]$ using $P[A = a \mid L = l]$, while standardization estimates $E[Y \mid A = a, L = l]$

3

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

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

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

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

[6] 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.

[7] 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.

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

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

[10] 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.