Abstract Title Page

Not included in page count.

Title: Beyond Binary: Using Propensity Scores to Account for Varying Levels of Program Participation in Randomized Controlled Trials

Author(s):
– Elizabeth A. Stuart, PhD, Department of Mental Health, Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health;
– Siri Warkentien, Doctoral Candidate, Department of Sociology, Johns Hopkins University;
– Booil Jo, PhD, Department of Psychiatry, Stanford University

Abstract Body

Limit 5 pages single spaced.

Background / Context:

Description of prior research and its intellectual context.

Propensity Scores

Originally proposed by Rosenbaum and Rubin (1983), propensity scores are a type of matching method that can be used with observational data to mimic a randomized experiment, at least with respect to the observed covariates (Stuart and Rubin, 2008). In a randomized experiment, participants are randomly assigned to either the treatment or control condition. The researcher can assume, based on the randomization, that the groups are balanced on all observed and unobserved characteristics prior to treatment. Any observed differences in outcomes following the treatment can then be attributed to the treatment itself, rather than to selection bias (Shadish, Cook, & Campbell, 2002). Propensity score matching uses this framework, constructing matched (or equated) treatment and control groups without including information on the outcome variable of interest (Stuart and Rubin, 2008). The two groups are as similar as possible based on a wide range of observed covariates that have been reduced to a single propensity score (Rubin, 1997), which is formally defined as the probability of receiving the treatment given those covariates. Ideally, the only difference between the treatment group and the control group in the propensity score analysis is that the treatment group actually received the treatment while the control did not. Then, similar to randomized experiments, the researcher can conclude that any difference post-treatment was caused by the treatment. A main benefit of using propensity score analysis is that it reduces selection bias due to observed covariates – one of the main threats to internal validity in quasi-experiments – because treatment and control groups are matched prior to treatment (Stuart & Rubin, 2008). Under the assumption that there is no unobserved confounding, unbiased treatment effect estimates can then be obtained using propensity score methods.

In the years since Rosenbaum and Rubin's groundbreaking work on causal inference through the use of propensity scores, researchers have applied it in many settings with available observational data. More recently, researchers have begun to consider applications for propensity scores that extend beyond the traditional binary treatment condition and into new settings. Imai and van Dyk (2004) extend the work of Joffe and Rosenbaum (1999) and Imbens (2000) to generalize the propensity score method to a "propensity function" to accommodate "arbitrary treatment regimes," including multiple dose levels, continuous levels, or multiple factors and their interactions. As one example, Imai and van Dyk used observational data to first estimate the effect of a continuous treatment that combines the frequency and duration of smoking by an individual. They then separate frequency and duration into two variables to present the propensity function with a bivariate treatment. In both instances, the data are subclassified based on the propensity function, with each subclass containing observations within a certain range of the propensity function. Results demonstrated that subclassification on the propensity function more effectively reduced bias and MSE compared to standard regression.

Randomized Trials

Randomized controlled trials (RCT) in education research are highly desirable because of their ability to make causal conclusions about the effect of the intervention under investigation. Yet, it is often the case that not all individuals assigned to the treatment comply with their assigned treatment. Furthermore, some individuals only partially comply with the assigned treatment. Under these conditions, the effect of the intervention may be obscured if the proportion of individuals assigned to treatment who are non- or partial-compliers is sufficiently large to mask the effects of the intervention on those who fully complied with treatment assignment. Researchers often account for this by estimating the complier average causal effect (CACE)—the effect of fully participating in the treatment—in addition to the effect of treatment group assignment, known as the intent to treat (ITT) effect (Stuart, Perry, Le, & Ialongo, 2009).

Propensity Scores in the RCT context

Recent work applying propensity scores in the context of RCTs has been undertaken by Jo and Stuart (2009). Expanding upon research by Follman (2000) and Joffe et al. (2003), Jo and Stuart use the framework of principal stratification to assess the practicality and performance of estimating "principal effects" using propensity score methods. Principal stratification refers to the classification of individuals based on potential values of intermediate variables, and forms the basis for the CACE described above. Often in the context of RCTs, the intermediate variables of particular interest are treatment receipt behavior, e.g., whether an individual is a complier or non-complier, where a complier is someone who would fully participate in the intervention when in the treatment group and would not participate in the control group. Because they are defined on the basis of behavior under both treatment and control conditions, the categories that result from principal stratification are not affected by treatment assignment; this permits calculation of principal effects – the treatment effect conditioned on the categories (the "principal strata"; Frangakis and Rubin, 2002). A complication even in a simple noncompliance setting with some assumptions applied is that stratum membership is known for individuals assigned to treatment, but unknown for those in the control condition (and sometimes it is not directly observed for anyone). Jo and Stuart (2009) combine the ideas of principal stratification and propensity scores to model compliance in the treatment group (for whom we observe their compliance status) and then use that model to estimate the probability of compliance for control group members. The treatment group compliers are then matched to the control group, finding the individuals in the control group who look like the compliers in the treatment group. In addition, they assess when covariate information obtained in the RCT of interest is sufficient for creating propensity scores to estimate principal effects, highlighting in part that strong predictors of compliance behavior are needed.

Purpose / Objective / Research Question / Focus of Study:

Description of the focus of the research.

The purpose of the current project is to explore the use of propensity scores to estimate the effects of interventions within randomized control trials, accounting for varying levels of implementation or fidelity. This work extends that of Jo and Stuart (2009) to settings with multiple or continuous measures of implementation. Rather than focus purely on a binary measure of compliance or non-compliance in the treatment group, the study looks at levels of compliance to treatment using a continuous measure of participation or fidelity. The methods are investigated using simulated data as well as data from the Infant Health and Development Program (IHDP), an eight-site randomized trial that targeted low-birth-weight premature infants, and offered comprehensive, quality early intervention care with the goal of reducing the developmental and health problems of low-birth-weight infants (IHDP, 1990).

Setting:
Description of the research location.
(May not be applicable for Methods submissions)

The Infant Health and Development Program is an eight-site randomized trial in which low-birth-weight, premature infants who were randomized to the treatment received early childhood development and family support services as well as pediatric follow-up care; infants who were randomized to the control received pediatric follow-up care only (IHDP, 1990). Services for the treatment group included home visits, child attendance at a high-quality center-based daycare during their second and third years, and parent group meetings. The initial primary outcome measures were collected when the participants were 36 months old using the Peabody Picture Vocabulary Test—Revised (PPVT-R) and the Stanford Binet Intelligence Scale.

Population / Participants / Subjects:
Description of the participants in the study: who, how many, key features or characteristics.
(May not be applicable for Methods submissions)

We use a similar subset of data as examined by Hill, Brooks-Gunn, and Waldfogel (2003). The sample consists of 985 infants who were members of the primary analysis group. Participants in this study had to meet the following qualifications: the infant was low-birth-weight (weigh less than or equal to 2500g); was premature (born at or before 37 weeks gestational age); lived within 45 minutes of the childcare center; and survived neonatal hospitalization (IHDP, 1990).

Intervention / Program / Practice:
Description of the intervention, program or practice, including details of administration and duration.
(May not be applicable for Methods submissions)

The goal of the Infant Health and Development Program intervention was to reduce the cognitive, behavioral, developmental, and health problems among low-birth-weight premature infants (IHDP, 1990). The intervention included two years of high quality center-based care at an early childhood development center during the child's second and third years of life. Although there were approximately 500 total possible days that children could attend the center, records were kept on daily attendance for each participant and large variation in the actual number of days attended resulted (Hill et al. 2003). Many outcome measures were collected on the children; however, we focus on the cognitive outcomes measured at age 3.

Significance / Novelty of study:
Description of what is missing in previous work and the contribution the study makes.

The significance of this study lies in its bridging of two areas of recent advances in the propensity score literature: expanding binary treatment into continuous measures of "treatment" and using propensity score methods to estimate quantities such as the CACE. As we briefly outlined in the Background section, on the one hand, researchers such as Joffe and Rosenbaum (1999), Imbens (2000), and van Dyk (2004) have advanced the previous binary "treatment" into multi-dose, bivariate, and continuous treatment measures. On the other hand, propensity scores have recently been applied to the randomized experiment setting (Jo & Stuart, 2009) for principal effect estimation. This study connects these two areas for the first time, allowing estimation of complex quantities.

We note that other methods for estimating CACE with continuous measures of treatment exist in the literature, including traditional instrumental variable (IV) models in econometrics (Angrist, Imbens & Rubin, 1996). Hill (2010) proposes the use of Bayesian Additive Regression Trees (BART) to investigate dosage effects, illustrated using the IHDP data. However, propensity scores offer an alternative approach that may serve as a sensitivity analysis to these other methods, as they rely on different assumptions. Relatively little research on the relative performance of both approaches under various conditions has been undertaken (Jo & Stuart, 2009).

Statistical, Measurement, or Econometric Model:

Description of the proposed new methods or novel applications of existing methods.

In this study, we use the generalized propensity score to model levels of compliance in the treatment condition. In particular, the propensity score models the number of days of center attendance by the child over the two-year period. This is an important advance over the traditional approaches, which have used a simple binary characterization of compliance, generally just measuring whether a child attended for more than 350 (or more than 400) days (Hill et al., 2003). With a continuous compliance measure we employ a linear model to estimate these values. This model is fit using the treatment group, for whom we have the number of days actually attended, and then the resulting model fit is used to predict levels of compliance (the generalized propensity score) for all children in the control group. We then subclassify treated and control group members by their predicted levels of compliance; within each subclass treated and control group members have similar levels of predicted compliance, allowing the comparison of treated individuals at a particular compliance level with controls who look as if they would have had that level of compliance had they been in the treatment group. By the properties of the generalized propensity score the baseline characteristics of the treated and control individuals within each subclass are also similar. Effects of varying levels of compliance are then estimated by comparing treatment and control group outcomes within each subclass.

Usefulness / Applicability of Method:

Demonstration of the usefulness of the proposed methods using hypothetical or real data.

In the current study, we demonstrate the usefulness of the proposed method using empirical data from a randomized controlled trial of an intervention strategy for low-birth-weight premature infants. In this particular study, the continuous measure of treatment compliance was number of days of childhood education center attendance. As the prevalence and funding of RCTs continues to increase and the importance of strong methodology for the estimation of causal effects remains high in education research, there will be numerous opportunities to use estimates for continuous measures of fidelity and implementation. These questions may be particularly important for researchers interested in getting into the “black box” and examining not just whether assignment to a particular treatment made a difference, but whether actual levels of participation changed outcomes. This study provides an alternative estimation approach for use in these settings and will be of broad use in education research.

Research Design:

Description of research design (e.g., qualitative case study, quasi-experimental design, secondary analysis, analytic essay, randomized field trial).
(May not be applicable for Methods submissions)

The study presents a combination of simulation studies and empirical investigation using the IHDP intervention. The simulation studies use data generating processes informed by the IHDP data, making the simulated data as realistic as possible. In particular, simulations use the observed samples and covariates from the IHDP sample, but with simulated outcome values such that the true dose/response relationships are known. This allows the assessment of the bias, mean square error, and coverage rates of alternative methods for estimating the effects of varying levels of compliance. We explore a range of outcome models, including linear and non-linear functions, and with varying strengths of the relationship between covariates and compliance levels. Simulations are done in R and Mplus, in a way similar to that described in Jo and Stuart (2009).

Data Collection and Analysis:

Description of the methods for collecting and analyzing data.
(May not be applicable for Methods submissions)

N/A

Findings / Results:

Description of the main findings with specific details.
(May not be applicable for Methods submissions)

N/A

Conclusions:

Description of conclusions, recommendations, and limitations based on findings.

This study shows the value in using the propensity score approach to estimate quantities such as the complier average causal effect. We demonstrate its usefulness in an empirical example using the Infant Health and Development Program data and highlight potential future uses for the approach. The study also reveals that the approach requires strong predictors of compliance level to work well. Therefore, we discuss limitations of the approach when strong predictors of compliance levels are not present in the data and advocate that researchers collect a broad array of data on study participants, especially variables that may be related to compliance level.

Appendices

Not included in page count.

Appendix A. References

References are to be in APA version 6 format.

Angrist, J. D., Imbens, G. W., & Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association, 91*, 444-455.

Follman, D. A. (2000). On the effect of treatment among would-be treatment compliers: an analysis of the Multiple Risk Factor Intervention Trial. *Journal of the American Statistical Association, 95,* 1101-1109.

Frankgakis, C. E. & Rubin, D. B. (2002). Principal stratification in causal inference. *Biometrics, 58,* 21-29.

Hill, J., Brooks-Gunn, J., and Waldfogel, J. (2003). Sustained effects of high participation in an early intervention for low-birth-weight premature infants. *Developmental Psychology, 39*(4): 730-744.

Hill, J. (2010). Bayesian Nonparametric Modeling for Causal Inference. *Journal of Computational and Graphical Statistics.* Epub ahead of print.

Imai, K. & van Dyk, D. A. (2004). Causal inference with general treatment regimes: Generalizing the propensity score. *Journal of the American Statistical Association, 99*(467), 854-866.

Imbens, G. W. (2000). The role of the propensity score in estimating dose-response functions. *Biometrika, 87,* 706-710.

Infant Health and Development Program. (1990). Enhancing the outcomes of low-birth-weight, premature infants: A multisite, randomized trial. *Journal of the American Medical Association, 263,* 3035-3042.

Jo, B. (2002). Estimation of intervention effects with noncompliance: Alternative model specifications. *Journal of Educational and Behavioral Statistics, 27*(4), 385-409.

Jo, B. & Stuart, E. A. (2009). On the use of propensity scores in principal causal effect estimation. *Statistics in Medicine, 28*, 2857-2875.

Joffe, M. M. & Rosenbaum, P. R. (1999). Propensity scores. *American Journal of Epidemiology, 150,* 327-333.

Joffe, M. M., The Have, T. R., & Brensinger, C. (2003). The compliance score as a regressor in randomized trials. *Biostatistics, 4,* 327-340.

Rosenbaum, P. R. & Rubin, D. B. (1983). The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika, 70,* 41-55.

Rubin, Donald. 1997. *Annals of Internal Medicine,127*(8), part 2.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference.* Boston: Houghton Mifflin.

Stuart, E. A. & Rubin, D. (2008). Best Practices in Quasi-Experimental Designs: Matching Methods for Causal Inference. Chapter in *Best Practices in Quantitative Methods*. J. Osborne, Editor. Thousand Oaks, CA: Sage Publications.

Stuart, E. A., Perry, D. F., Le, H. N., & Ialongo, N. S. (2008). Estimating intervention effects of prevention programs: Accounting for noncompliance. *Prevention Science, 9,* 288-298.

Appendix B. Tables and Figures

Not included in page count.