SREE 2018 Symposium

Valid Causal Inference in Single-Site and Multi-Site Studies: Lessons for Analysts

Organizer: Guanglei Hong (University of Chicago)

Discussant: Luke Miratrix (Harvard University)

Overview

This symposium will share new methodological developments useful for enhancing research evidence in single-site and multi-site program evaluations. Even though an increasing number of studies on educational effectiveness have employed theoretically sound research designs involving randomized treatment assignment that is sometimes combined with random sampling from a well-defined population, the internal validity and the external validity of the analytic results are not guaranteed. We present three inter-related studies that highlight the potential pitfalls and offer new solutions to data analysts.

Even when a multi-site randomized controlled trial is perfectly implemented, a correlation between the site-specific impact and its estimation precision threatens to bias the estimation of the average treatment effect (ATE). Suboptimal analysis of the data may inadvertently either introduce bias or increase noise. The first study proposes an optimal solution to this bias-variance trade-off problem in multisite analysis.

In the context of analyzing a national multi-site randomized trial that employed a complex sample and survey design and suffered from non-random attrition, the second study clarifies the logic of incorporating sample weights and estimating non-response weights that are necessary for ensuring the external and internal validities of the ATE estimation. These weights are then combined with ratio-of-mediator probability weighting for causal mediation analysis. The study also addresses challenges associated with specification error and estimation error that arise in a series of propensity score-based weighting.

The third study focuses on a weighting-based approach to sensitivity analysis. In experimental or non-experimental research, all propensity score-based weighting strategies share the limitation of not being able to handle omitted confounding. Sensitivity analysis is essential for quantifying the remaining selection bias and assessing its consequence. The weighting-based approach overcomes some major limitations of conventional model-based approaches to sensitivity analysis. It is broadly applicable to evaluations of the average treatment effect on the treated (ATT), the ATE, as well as the direct effect and the indirect effect in single-site and multi-site studies.

We believe that many applied researchers in education can benefit from these new developments in methodological work. To make the technical results accessible to a broad audience, all three presentations and the discussion will highlight big conceptual ideas, logical intuitions, and key take-home messages. The presenters will make use of pedagogical examples, will share simulation results and real data applications, and will place emphasis on developing tools for routine use by data analysts.

1

Estimating the Average Treatment Effect in Multi-Site Randomized Trials When Sample Sizes Are Endogenous and Treatment Effects Vary

Stephen W. Raudenbush
Daniel Schwartz
University of Chicago

In a multi-site randomized trial, sites such as classrooms, schools or districts are sampled; within each site, units are assigned at random to treatments. The US Institute for Education Sciences has funded hundreds of large-scale multi-site randomized trials (Spybrook, 2013). Our aim is to compare standard statistical methods for estimating the average treatment effect in such studies. We find that conventional regression methods using fixed or random effects can produce biased estimates. These methods weight site-specific impact estimates by estimates of their precision, but these precisions can be correlated with site-specific impacts. Introducing design weights will remove such bias but may substantially increase standard errors. To address this bias-variance tradeoff, we introduce a new estimator that, for studies with many sites, is more efficient than conventional estimators using precision-weights or design weights.

We first use potential outcomes and a super-population framework to precisely describe different target populations and therefore different definitions of the average treatment effect in such multi-site trials. In a two-level study, for example, the sites may be schools and students may be assigned at random to treatments within schools. The experimenter may wish to generalize results to a population of students or to a population of schools. If the number of students who constitute the sub-population of each school varies across schools, the population average treatment effect will tend to differ for these two target populations whenever site-specific treatment effects are heterogeneous. In a three-level study, we may have students nested within teachers within schools. In many such studies the teachers within each school will be assigned at random to treatments. In this case, one can construct a population of students, a population of teachers, or a population of schools, depending on the substantive aims of the study. An average treatment effect defined on each of these populations will tend to differ when impacts are heterogeneous.

Second, we argue that site-specific sample sizes and probabilities of treatment assignment will tend to be endogenous in many multi-site trials in education. Many of these studies use a lottery to assign students to treatments. For example, in charter school lottery studies, parents within a given site apply for seats in the charter school. When the number of applicants exceeds the number of available seats in the charter school, offers of admission are based on a lottery. However, the number of applicants and the number of seats available may depend on the popularity of the school as well as the number of alternative schooling options locally available. As a result, the sample size and probability of admission may be related to the impact of the charter school. In this setting, standard methods of estimation such as ordinary least squares with site fixed effects or generalized least squares with random coefficient models will tend to introduce bias because these methods weight site-specific estimates of impact by precisions that depend on the sample size and the probability of assignment to treatment. Specifically, sites that have larger sample sizes and propensity scores near 0.50 will have greater precision and therefor will receive a larger weight than will sites with smaller sample sizes and propensities farther from 0.50. Suppose, for example, the aim is to generalize to a population of sites. If sites with more applicants are more effective, precision weighting will over-estimate the average treatment effect.

We'll show how introducing design weights removes the bias. The weights are of two types: inverse-probability-of-treatment weights (Robins, Hernan, & Brumback, 2000) enable us to simulate a study in which every site has the same sample size; Horvitz-Thompson weights ensure that the sampled units represent the desired population (Horvitz and Thompson, 1952). However such weighting may produce embarrassingly inefficient estimates. In particular, sites having imprecise impact estimates may be assigned large weight, so that the unbiased estimator may be very imprecise.

Our illustrative example uses data from the National Head Start Impact Study (Puma et al., 2010). In this study, 317 Head Start centers were selected at random from a list of all such centers in the US. Within each site, the experimenters sought applicants for available slots in the local Head Start program. An attractive target population in this setting is the universe of all Head Start centers. However, the number of students applying to Head Start and the probability of winning the lottery varied dramatically across sites. We'll show that the unbiased estimator using design weights is so inefficient that its standard error can be substantially reduced by throwing out the data from many small sites.

Thus a challenging bias-variance tradeoff emerged: conventional precision-weighted estimators reduce variance by assigning a larger weight to sites that produce more reliable impact estimates. But these estimators will be biased even in large samples if site-specific sample sizes are related to site-specific impacts. In contrast, estimators that use design weights to remove bias may be unacceptably noisy.

Our third aim, therefore, is to seek an approach that eliminates bias while producing acceptably small variance. We shall introduce a novel estimator that adjusts for the natural logarithm of the site-specific sampling precision and includes site-specific random coefficients to represent heterogeneity of treatment effect. We prove that, under the assumption that site-specific impacts and site-specific log-precisions are bivariate normal in distribution, this estimator is asymptotically more efficient than either of the standard precision-weighted estimators or design-weighted estimators. Specifically, adjusting for the logarithm of site-specific precision eliminates bias but allows sites with large precision to contribute somewhat more to the overall average than do sites with less precision. Fourth, we use a simulation study to evaluate the efficiency of this approach in finite samples and when parametric assumptions do not hold. We illustrate these ideas using data from the National Head Start Impact Study (Puma et al., 2010).

References

Horvitz. G., and Thompson, D.J. (1952). A Generalization of Sampling without Replacement from a Finite Universe. *Journal of the American Statistical Association, 47*(260), 663-685.

Puma, M., Bell, S. Cook, R., Heid, C., Shapiro, G., Broene, P., Jenkins, F., Fletcher, P., Quinn, L., Friedman, J., Ciarico, J.,Rohacek, M., Adams, G., Spier, E. (2010). *Head Start Impact Study. Final Report.* US Administration for Children & Families.

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

Spybrook, J. (2013) “Detecting Intervention Effects Across Context: An Examination of the Precision of Cluster Randomized Trials,” *The Journal of Experimental Education*, DOI: 10.1080/00220973.2013.813364.

2

A Template for Multi-site Causal Mediation Analysis

Xu Qin
Guanglei Hong
University of Chicago

Jonah Deutsch
Mathematica Policy Research

Edward Bein
U.S. Food and Drug Administration

Context

In a large-scale multi-site randomized trial, the sample is sometimes chosen deliberately to represent the target population. Due to treatment randomization within each site, program impacts can be identified site by site. Such a design allows analysts to estimate, without bias, the average program impact and its between-site variance that are generalizable to the population of individuals or sites. However, the generalizability would be compromised if analysts pay little attention to the complex sample and survey design; while non-random attrition in longitudinal follow-ups poses a familiar threat to the internal validity of the causal conclusions.

To test a program theory explaining why the program produces its intended effect, researchers may decompose the average program impact into an indirect effect transmitted through a hypothesized focal mediator and a direct effect attributable to all other possible pathways. By examining the between-site heterogeneity of these causal effects, researchers may further evaluate the generalizability of the program theory. Additional challenges arise, however, because the mediator is typically not randomized and hence the mediator-outcome relationship is likely confounded by selection factors.

Objectives

In this study, we present a template for multi-site causal mediation analysis that addresses the above challenges. We explain to data analysts the logic of incorporating sample weights suitable for a given complex sample and survey design, of employing non-response weights to account for non-random attrition, and of combining the above two weights with ratio-of-mediator-probability weighting (RMPW) to unpack causal mechanisms. We clarify the identification assumptions under which the analytic results are externally and internally valid. Because the propensity score models for response and those for the mediator must be specified by the researcher and analyzed with the sample data, model specification error and estimation error become additional concerns. We develop analytic procedures that empirically detect specification error through a multi-step examination of observed covariate balance. A series of weighting-based sensitivity analyses further assess the consequences of potential omitted confounders. In addition, we derive the asymptotic variance for the weighted estimators of the causal effects, taking into consideration estimation errors in all the propensity score analyses. We implement this template in a re-analysis of data from the National Job Corps Study (NJCS).

Application

NJCS is a multi-site randomized evaluation of the nation’s largest education and training program for disadvantaged youth ages 16 to 24. Job Corps program theory and past empirical research have suggested that helping students obtain an education credential is a primary pathway through which Job Corps promoted economic independence (Flores & Flores-Lagunes, 2013; Lee, 2009; Schochet, Burghardt, & McConnell, 2006, 2008). However, the program did not seem to produce the same economic benefit over all the sites (Johnson et al, 1999). Is it because the program mechanism mediated by educational attainment--the indirect effect--operated differently at different centers? Or is it because the roles of other program or non-program elements--summarized in a direct effect—varied across the sites? Such evidence will be crucial for enriching theoretical understanding and for informing the design and implementation of educational programs alike.

NJCS involved all the 100+ Job Corps centers in the nation. The sample universe included 80,883 eligible youths who applied for Job Corps between November 1994 and February 1996. 15,386 youths were selected into a nationally representative research sample at the baseline but had different probabilities of being included in the follow-up interview samples. All the sampled youths at the baseline were assigned at random to either the program group or the control group. Those in the control group were barred from enrolling in Job Corps for three years. Program group and control group members who were initially assigned to the same Job Corp center constitute the sample of individuals at the given site. The mediator, collected at the 30-month follow-up, indicates whether a youth had obtained an education credential since the randomization. The outcome is weekly earnings at the 48-month follow-up. However, some sampled youths were lost to attrition or failed to provide specific information on education or earnings.

We select a sample weight that accounts for the differential sampling probabilities at the 48-month follow-up. We then estimate each sampled individual's propensity of responding to the survey measures of the mediator and the outcome given the individual's treatment assignment, site membership, and baseline characteristics. The estimated non-response weight is inverse to the estimated propensity of response. Subsequently, we estimate each respondent's propensity of obtaining an education credential under each treatment condition. The estimated RMPW weight is a ratio of the estimated mediator propensity under the control condition to that under the experimental condition for each respondent in the program group; the RMPW weight is 1 for the respondents in the control group. Applying a product of all three weights, we estimate, for each site, (A) the average earnings if all the eligible youths had been assigned to the program group, (B) the average earnings if they had all been assigned to the control group, and (C) the average earnings if they had all been assigned to the program group yet had counterfactually received no program benefit in educational attainment. (A) – (C) estimates the site-specific indirect effect; and (C) – (B) estimates the site-specific direct effect. Aggregating these results over all the sites, we obtain estimates of the population average indirect effect and direct effect and their between-site variances and covariance.

Discussions

Multi-site trials employing complex sample and survey designs offer unique opportunities yet also pose multiple challenges in attempts to test program theories. In practice, there has been confusion among data analysts with regard to how to appropriately apply sample weights, how to construct non-response weights in light of the complex survey design, how to combine them with other propensity score-based weighting strategies in causal mediation analysis, how to assess the adequacy of each weighting adjustment, and how to quantify sampling uncertainty in a multi-step estimation problem. We use simulation results to reveal potential consequences of various missteps commonly seen in past research. We develop an open-source R package that provides applied researchers with a convenient implementation tool.

References

Flores, C. A., & Flores-Lagunes, A. (2013). Partial identification of local average treatment effects with an invalid instrument. *Journal of Business and Economic Statistics*, *31*(4), 534-545. Johnson, T., Gritz, M., Jackson, R., Burghardt, J., Boussy, C., Leonard, J., & Orians, C. (1999). *National Job Corps Study: Report on the process analysis*. Princeton, NJ: Mathematica.

Lee, D. S. (2009). Training, wages, and sample selection: Estimating sharp bounds on treatment effects. *Review of Economic Studies*, *76*(3), 1071-1102.

Schochet, P. Z., Burghardt, J., & McConnell, S. (2006). *National Job Corps Study and longer-term follow-up study*: *Impact and benefit-cost findings using survey and summary earnings data*. *Final Report*. Mathematica Policy Research, Inc.

Schochet, P. Z., Burghardt, J., & McConnell, S. (2008). Does Job Corps work? Impact findings from the National Job Corps Study. *American Economic Review*, *98*(5), 1864-1886.

3

Weighting-Based Approach to Sensitivity Analysis in Single- and Multi-Site Studies

Guanglei Hong
Xu Qin
University of Chicago

Fan Yang
University of Colorado Denver

Relevance

Causal inference relies on the assumption that selection bias has been eliminated through research design or statistical adjustment. When this assumption becomes overly strong due to limitations in the design or in the adjustment method, a sensitivity analysis becomes a necessary step in the evaluation (Frank, 2000; Rosenbaum, 1986, 2002). The goal is to determine whether a conclusion could be easily reversed by a plausible violation of the assumption. Causal conclusions that are harder to alter by such a violation are expected to add a higher value to scientific knowledge about causality.

Objective

This study develops a series of weighting-based strategies for quantifying potential bias in evaluations of the average treatment effect on the treated (ATT), the average treatment effect (ATE, equivalent to the intention-to-treat (ITT) effect when focusing on the effect of the treatment assignment), and the natural direct effect (NDE) and the natural indirect effect (NIE) that decompose the ATE. Because treatment impacts and mechanisms may depend on the settings (such as schools or communities), multisite studies help reveal between-site heterogeneity in these causal effects (Qin & Hong, 2017; Raudenbush & Bloom, 2015). The weighting-based approach to sensitivity analysis is coherent with a large class of propensity score-based weighting methods. In its essence, the discrepancy between a new weight that adjusts for an omitted confounder and an initial weight that omits the confounder captures the role of the confounder that contributes to the bias.

Single-Site Evaluations

ATT

In non-experimental studies, or in experimental studies that suffer from non-random attrition, selection bias becomes inevitable in evaluating the ATT defined as $E[Y(1) - Y(0)|Z = 1]$. Here Z is the treatment indicator taking value 1 if an individual is assigned to the experimental condition and 0 otherwise; $Y(1)$ and $Y(0)$ denote the corresponding potential outcomes. In an initial analysis that adjusts for a vector of observed pretreatment covariates denoted by X, $W_{0.ATT} = pr(Z = 1|X)/pr(Z = 0|X)$ applied to the control group transforms the pretreatment composition of the control group to resemble that of the experimental group. Under the assumption that the experimental group and the control group have the same distribution of $Y(0)$ within levels of $X = x$, $E[Y|Z = 1] - E[W_{0.ATT}Y|Z = 0]$ identifies the ATT.

However, if the identification assumption requires conditioning on P in addition to X, the ATT will be identified by $E[Y|Z = 1] - E[W_{0P.ATT}Y|Z = 0]$ instead, where $W_{0P.ATT} = pr(Z = 1|X, P)/pr(Z = 0|X, P)$. When P is omitted from the causal analysis, the bias can be quantified as the following:

$$Bias_{ATT} = \{E[Y|Z = 1] - E[W_{0.ATT}Y|Z = 0]\} - \{E[Y|Z = 1] - E[W_{0P.ATT}Y|Z = 0]\}$$

$$= cov(W_{0P.ATT} - W_{0.ATT}, Y|Z = 0).$$

The last equation holds because $E[W_{0.ATT}|Z = 0] = E[W_{0P.ATT}|Z = 0] = 1$. The effect size of $Bias_{ATT}$ is a product of two sensitivity parameters:

$$ES \text{ of } Bias_{ATT} = \sigma_{0.ATT}\rho_{0.ATT},$$

where $\sigma_{0.ATT} = \sqrt{var(W_{0P.ATT} - W_{0.ATT}|Z = 0)}$ is associated with the between-group difference in P; and $\rho_{0.ATT} = corr(W_{0P.ATT} - W_{0.ATT}, Y|Z = 0)$ is associated with the degree to which P predicts Y.

To illustrate, suppose that an initial estimate of the effect size of ATT is positive and is not statistically significant and that the 95% confidence interval is (-0.1, 0.5). If we remove a negative ES of bias $\sigma_{0.ATT}\rho_{0.ATT} < -0.1$, the new 95% CI would shift entirely to the positive side and would lead to a new conclusion that the ATT is positive. If we remove a positive ES of bias $\sigma_{0.ATT}\rho_{0.ATT} > 0.5$, the new 95% CI would shift entirely to the negative side and an opposite conclusion would be drawn. The observed pretreatment covariates in the current study and those from past research supply a range of plausible reference values for the two sensitivity parameters. Such information will enable the analyst to determine, for example, that an ES of bias close to but less than -0.1 is plausible whereas an ES of bias greater than 0.5 is nearly implausible. In this case, the results would suggest that the null finding obtained in the initial analysis is sensitive to potential (or actual) omitted confounders. Moreover, the analyst would conclude that, while evidence for a negative ATT is absent, a positive ATT cannot be ruled out.

ATE

ATE is defined as $E[Y(1) - Y(0)]$. In an initial analysis, inverse-probability-of-treatment weighting (IPTW) transforms the joint distribution of X in the experimental group and the control group to resemble the distribution in the entire population (Robins, 2000; Rosenbaum, 1987). Non-response weighting serves the same purpose. Considering an identification that requires additional adjustment for P, we similarly derive the bias due to the omission of P and represent its effect size as a function of sensitivity parameters.

NIE and NDE

A causal mediation analysis in its simplest form evaluate the treatment effect on the outcome transmitted through a focal mediator M, defined as $NIE = E\left[Y\left(1, M(1)\right) - Y\left(1, M(0)\right)\right]$, and the treatment effect transmitted through other unspecified mechanisms, defined as $NDE = E\left[Y\left(1, M(0)\right) - Y\left(0, M(0)\right)\right]$. When the treatment is randomized, the NIE and the NDE are identified under the strong assumptions that there is no omission of pretreatment or posttreatment confounders of the mediator-outcome relationship. Ratio-of-mediator-probability weighting (RMPW) transforms, within levels of X, the mediator distribution in the experimental group to resemble that in the control group (Hong, 2010, 2015; Hong, Deutsch, & Hill, 2015). Our forthcoming article in JEBS develops a weighting approach to assessing bias in NIE and that in NDE due to omitted pretreatment or posttreatment confounders.

Multi-Site Evaluations

In multi-site studies, the causal parameters of interest include not only the average ATT, ATE, NIE, and NDE, but also the between-site variance of each. Bias due to omitted confounders may sneak into not only the estimated average effects but also the estimated between-site variance and covariance. This is because an omitted confounder may not operate the same in all the sites. The weighting-based approach can be applied within each site. We then aggregate the site-specific bias over all the sites.

Contributions

We will show that, in single-site and multi-site studies, the new approach is considerably more flexible and promise broader applicability than most conventional sensitivity analysis strategies that tend to be constrained by a linear, additive framework. In addition to relaxing a host of often unrealistic parametric assumptions, the new approach makes it possible to assess the aggregate bias associated with multiple omitted covariates. We have developed tabular and graphical forms to display sensitivity analysis results and will supply R code to ease the computation.

References

Frank, K. (2000). Impact of a confounding variable on the inference off a regression coefficient. *Sociological Methods and Research*, *29*(2), 147-194.

Hong, G. (2010). Ratio of mediator probability weighting for estimating natural direct and indirect effects. *Proceedings of the American Statistical Association, Biometrics Section* (pp. 2401–2415). Alexandria, VA: American Statistical Association.

Hong, G. (2015). *Causality in a social world: Moderation, mediation and spill-over*. West Sussex, England: John Wiley.

Hong, G., Deutsch, J., & Hill, H. D. (2015). Ratio-of-mediator-probability weighting for causal mediation analysis in the presence of treatment-by-mediator interaction. *Journal of Educational and Behavioral Statistics, 40*, 307–340.

Qin, X., & Hong, G. (2017). A weighting method for assessing between-site heterogeneity in causal mediation mechanism. *Journal of Educational and Behavioral Statistics, 42*(3), 308-340.

Raudenbush, S. W., & Bloom, H. (2015). Using multi-site randomized trials to learn about and from a distribution of program impacts. *American Journal of Evaluation*, *36*, 475–499.

Robins, J. M. (2000). Marginal structural models versus structural nested models as tools for causal inference. In M. E. Halloran & D. Berry (Eds.), *Statistical Models in Epidemiology, the Environment, and Clinical Trials* (pp. 95–133). New York, NY: Springer.

Rosenbaum, P. R. (1986). Dropping out of high school in the United States: an observation al study. *Journal of Educational Statistics*, *11*, 207-224.

Rosenbaum, P. R. (1987). Model-based direct adjustment. *Journal of the American Statistical Association*, *82*, 387–394.

Rosenbaum, P. R. (2002). *Observational Studies* (2nd ed.). NY: Springer.