Title:
Using covariates to detect treatment effect heterogeneity in multisite trials

Author:
Luke W. Miratrix
Harvard Graduate School of Education
lmiratrix@g.harvard.edu
(Presenting author)

Using covariates to detect treatment effect heterogeneity in multisite trials

Background / Context: Treatment effect heterogeneity is a critical component in understanding the results of large-scale randomized trials. For overview discussions see Schochet, Puma, & Deke (2014) or Weiss, Bloom, & Brock (2014). For example, in the project motivating this work, researchers are investigating contextual effects on an intervention to improve preschool classroom learning environments (as part of a secondary analysis of the National Center for Research on Early Childhood Education Professional Development Study (NCRECE PDS)). In these exploratory analyses, it is important to control testing procedures while maintaining high levels of power. One way forward is to first conduct a "gateway" test of whether there is substantial treatment effect variation before proceeding with systematic exploration.

The question is then how to conduct such an omnibus test in a maximally powerful way. Current methods (e.g., Bloom et al., 2013, Raudenbush & Bloom, 2015, Weiss et al., 2016) look for evidence of variation across site but do not take advantage of any site level covariates that may predict such variation. These approaches could be considered tests for *idiosyncratic variation*, variation not explicitly tied to covariates, as compared to testing for *systematic variation*, variation explicitly modeled as a function of covariates (see Heckman et al., 1997, or Djebbari and Smith, 2008). The question is then how to best take advantage of a variable believed to, at least in part, predict site level variation. Ideally, exploiting such covariates could answer Bloom and Spybrook (2017), which finds that one typically needs quite large multisite trials to detect cross-site variation.

Purpose / Objective / Research Question: We have two primary methodological research questions: (1) How does one best demonstrate the existence of cross-site treatment effect variation in a multisite trial if one has covariates predictive of such variation? and (2) What are the tradeoffs in using covariates for doing this?

Classic methods either rely on ANOVA-style variance calculations (i.e., with Q-statistics from meta-analysis, see Weiss et al., 2016) or test for the presence of an interaction term between a given site-level covariate and treatment in a linear model. See Appendix. In some contexts, such as with a moderately predictive covariate, either of these two approaches (within a multilevel modeling framework) may not be optimal. We propose to therefore use a hybrid test that tests for both systematic and idiosyncratic variation simultaneously using an adjusted likelihood ratio test.

Research Design: The main contributions of this investigation are, first, an overview of different methods one might use to detect treatment variation, and, second, a principled multifactor simulation experiment to systematically compare the performance of these methods. The simulation has two major arms to assess validity and power. We primarily examine

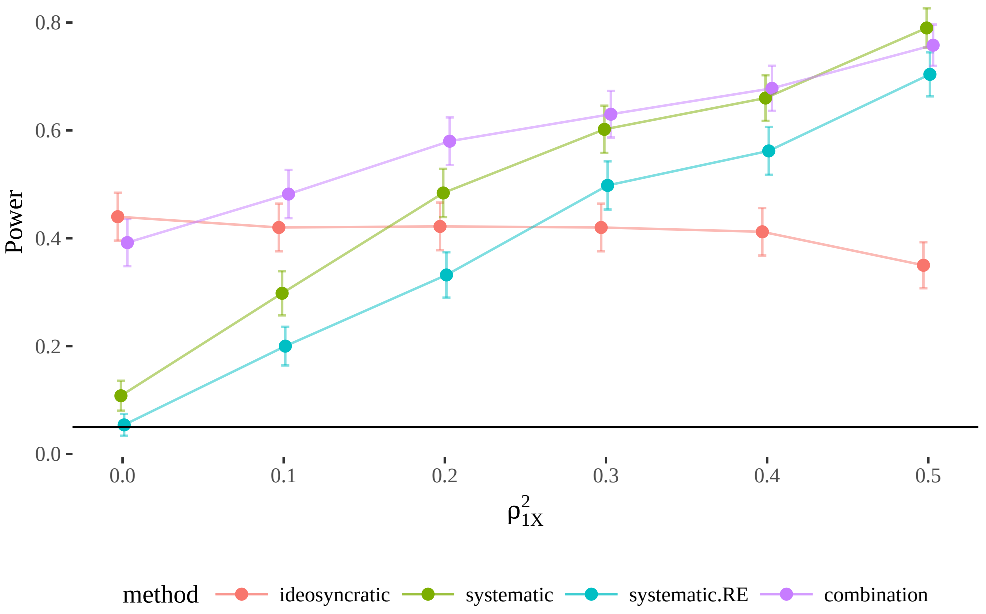

Figure 1: Power vs. predictive power of covariate for 4 methods for detecting cross-site treatment heterogeneity. Simulation setting has $J = 30$ sites with average size of $\bar{n} = 20$ teachers per site. Overall variation is 0.20. Idiosyncratic methods superior for low predictive covariates, systematic tests for high. Hybrid methods generally outperform either.

how different methods are more or less powerful depending on the predictiveness of the site level covariates. Primary factors we systematically vary are (1) amount of underlying site level variation, (2) number and size of sites, (3) the predictive power of covariates for treatment effect variation. We control other aspects of the data generating model to preserve the marginal variance of control and treatment outcomes in order to maintain an effect size interpretation to all results. See Appendix for further detail.

Initial Findings: As one part of our results, Figure 1 shows a simulation setting with a fixed cross site standard deviation of average effects of 0.20. We here examine the three naïve multilevel modeling approaches and the likelihood ratio hybrid approach. Under no variation, these methods had rejection rates of approximately 0.04; i.e., all appeared to be valid or even slightly conservative tests in this context. As expected, as the predictive power of the site-level covariate increases, the power of the systematic tests also increases. The idiosyncratic test (testing for *total* variation without including the covariate in the model) is essentially stable. Finally, the hybrid tests are generally more powerful than both the other approaches if the covariate is mildly to moderately predictive.

We have also identified some potentially surprising aspects of these testing approaches. For example, depending on the model specification used for testing for systematic effect, one can have elevated rejection rates if there is idiosyncratic heterogeneity (see far left of Figure 1). Depending on interpretation, this could be considered as either an invalid testing procedure, or a valid, but low-power, testing procedure of overall variation.

Conclusions: The hybrid test does have some cost when the covariates predicting treatment variation are either quite weak or quite strong. That being said, when one believes one has a reasonably predictive covariate, it should be incorporated as part of a hybrid test for variation.

Next steps are to develop this overall finding in greater detail, and to extend the overall intuition from the multilevel modeling framework to the more model independent framework of detecting systematic and idiosyncratic variation using methods based on the assignment mechanism, such as presented in Ding et al. (2016). In particular, one advantage of the Q-statistic approach is it does not make major distributional or modeling assumptions, as compared to multilevel modeling. But it also does not allow for easy incorporation of covariates. We extend this approach with an initial modeling step of estimating the site-level effects in order to increase power. We will also apply the methods that perform best to the NCRECE trial data.

Bibliography

Bloom, H. S., & Spybrook, J. (In Press). Assessing the Precision of Multisite Trials for Estimating the Parameters of a Cross-Site Population Distribution of Program Effects. *Journal of Research on Educational Effectiveness.*

Ding, Peng, Avi Feller, & Luke Miratrix. (2016) Decomposing treatment effect variation. arXiv preprint arXiv:1605.06566.

Djebbari, H., & Smith, J. (2008). Heterogeneous impacts in PROGRESA. *Journal of Econometrics*, 145(1-2), 6480.

Heckman, J. J., Smith, J., & Clements, N. (1997). Making the most out of programme evaluations and social experiments: Accounting for heterogeneity in programme impacts. *The Review of Economic Studies*, 64(4), 487535.

Schochet, P. Z., Puma, M., & Deke, J. (2014). Understanding Variation in Treatment Effects in Education Impact Evaluations: An Overview of Quantitative Methods. NCEE 2014-4017. *National Center for Education Evaluation and Regional Assistance.*

Snijders, T., & Bosker, R. (2004). *Multilevel analysis: An introduction to basic and applied multilevel analysis, 2nd Ed.* London: Sage.

Weiss, M. J., Bloom, H. S., & Brock, T. (2014). A conceptual framework for studying the sources of variation in program effects. *Journal of Policy Analysis and Management*, 33(3), 778-808.

Weiss, M. J., Bloom, H. S., Savitz, N. V., Gupta, H., Vigil, A., & Cullinan, D. (In Press). How Much Do the Effects of Education and Training Programs Vary Across Sites? Evidence from Existing Multisite Randomized Control Trials. *Journal of Research on Educational Effectiveness.*

Appendix

Classic multilevel model for estimating treatment variability

Consider a multi-site trial with K sites and n_j teachers in each site. Let X_j be a site-level covariate, such as neighborhood context, that is thought to predict both general teacher outcome and treatment impact (e.g., if X_j is site climate we might imagine both general outcomes and treatment impact to be higher for supportive sites). The classic multilevel model would then be, for teacher i in site j:

$$\begin{aligned} Y_{ij} &= \beta_{0j} + \beta_{1j} Z_{ij} + \epsilon_{ij} \\ \beta_{0j} &= \gamma_{00} + \gamma_{01} X_j + u_{0j} \\ \beta_{1j} &= \gamma_{10} + \gamma_{11} X_j + u_{1j} \end{aligned}$$

with

$$\begin{pmatrix} u_{0j} \\ u_{1j} \end{pmatrix} \sim N \left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \ \begin{pmatrix} \tau_{00} & \tau_{01} \\ & \tau_{11} \end{pmatrix} \right]$$

and $\epsilon_{ij} \sim N(0, \sigma_\epsilon^2)$.

This is a random-slope model, with β_{0j} being the average teacher outcome (under control) for site j, and β_{1j} being the average treatment effect on teachers in site j. Much recent literature proposes instead using a model with a fixed effect for the β_{0j} to account for differering proportions of treatment and other concerns; we will extend our simulations to this context.

We are interested in detecting treatment effect heterogeneity, i.e., whether site-average treatment impacts differ at all (across sites). Heterogeneity can arise from two things: a systematic effect from X_j ($\gamma_{11} \neq 0$), and an idiosyncratic effect ($\tau_{11} > 0$). If we ignore X_j then, if we assume the distribution of X_j across sites has some variance σ_X^2, our total treatment heterogeneity is

$$\tau_{11}^* = \text{var}\{\beta_{1j}\} = \gamma_{11}^2 \text{var}\{X_{ij}\} + \text{var}\{u_{ij}\} = \gamma_{11}^2 \sigma_X^2 + \tau_{11}.$$

Therefore, the proportion of treatment effect variation explained by X is ρ_{1X}^2:

$$\rho_{1X}^2 = \frac{\gamma_{11}^2 \sigma_X^2}{\tau_{11} + \gamma_{11}^2 \sigma_X^2}.$$

This is the R_τ^2 measure from Ding et al. (2016).

We are interested in exploiting X_j to detect overall treatment variation. We can do this by testing for γ_{11} being non-zero (a systematic test), testing for τ_{11} being non-zero (a parametric version of the Q-statistic method described below), or, in the hybrid case, by testing for both γ_{11} *and* τ_{11} being zero simultaneously. Other models are possible, e.g., an idiosyncratic test not including the interaction of X_j and Z_{ij} (which is what is displayed on Figure 1).

The naïve ideosyncratic tests do not take full advantage of the necessarily positive variance parameters. Following, e.g., Snijders and Bosker (2004) we can compare change in deviance from our null model and our full model allowing for both idiosyncratic and systematic variation to a "chi-bar" distribution to achieve greater power (the naïve tests are conservative.) This extension is also necessary for the hybrid approach.

The core simulation study framework

We here describe how we generate data for our simulation study, including how to set the various parameters in terms of more directly interpretable parameters. We index our Data Generating Process (DGP) with the following parameters:

1. $\bar{n}$, the expected sample size per site,
2. K, the number of sites.
3. p, the proportion of units treated (roughly the same across sites),
4. τ^*_{11}, the total treatment effect variability,
5. ρ^2_{0X}, the proportion of the site control mean outcome variability explained by X,
6. ρ^2_{1X}, the proportion of the treatment variability explained by X,
7. ICC, the proportion of (control-side) variability explained by site, and
8. γ_{00}, γ_{10}, the average outcome and average treatment effect (these are nuisance parameters in this context; we set them to 0 and 0.2).

We generate plausible data as follows. First, we generate site covariates and characteristics (size):

$$\begin{aligned} n_j &= Poisson(\bar{n}) \\ X_j &= N(0,1) \end{aligned}$$

For the Z_{ij}, we, mimicking the actual design of NCRECE PDS, conduct a complete randomization, *ignoring site of the individuals*, by assigning pn teachers to treatment, with $n = \sum n_j$ (instead of a randomization within each block).

We then generate outcomes using our original full model, setting the free parameters as follows (these depend on the assumption that the variance of the $Y_i(0)$ across sites is 1, i.e., that we are dealing with effect size units):

$$\begin{aligned}
\gamma_{01} &= \rho_{0X}\sqrt{ICC}/\sigma_X \\
\gamma_{11} &= \rho_{1X}\sqrt{\tau^*_{11}} \\
\tau_{00} &= (1-\rho^2_{0X})ICC \\
\tau_{01} &= -\tau^*_{11}/2 \\
\tau_{11} &= (1-\rho^2_{1X})\tau^*_{11} \\
\sigma^2_\epsilon &= 1 - ICC
\end{aligned}$$

Derivations, not shown, show that these settings achieve the desired ICC, etc.

Commentary. We generate data so the marginal variances do not blow up in the presence of heterogeniety. In particular, the above allows us to manipulate structure without changing overall treatment variation or the marginal variances of the control or treatment units. We could instead let the treatment variance go up with treatment variability by relaxing constraints on the τ_{01} term. For example, we could set $\tau_{01} = 0$.

The above uses a continuous X_j. We also consider dichotomous X_j, such as a site-level dichotomized climate variable (where 1 indicates the site supported the program, and 0 that the site ignored it). The above formula generally follow, but with additional σ^2_X terms.

The Q-statistic test for idiosyncratic variation

Following Weiss et. all (2016) the Q-statistic is

$$Q = \sum_{j=1}^{J} \frac{(\hat{\beta}_{1j} - \bar{\beta})^2}{v_j}$$

where v_j are the estimated precisions of the β_{1j}s and $\bar{\beta}$ is a precision-weighted average of the estimated $\hat{\beta}_{1j}$.

Under the null of no variation, Q has an approximate χ^2 distribution with $J-1$ degrees of freedom, allowing one to test for heterogeneity.

We extend this by estimating the β_{1j} using covariates. We estimate uncertainty with the bootstrap, as the χ^2 approximation may no longer hold here.