Abstract Title Page

Not included in page count.

Title: Methods for modeling and decomposing treatment effect variation in large-scale randomized trials

Authors and Affiliations:
Peng Ding, Harvard University
Avi Feller, Harvard University
Luke Miratrix, Harvard University

Abstract Body

Limit 4 pages single-spaced.

Background / Context:

Description of prior research and its intellectual context.

Recent literature has underscored the critical role of treatment effect variation in estimating and understanding causal effects. This approach, however, is in contrast to much of the foundational research on causal inference. Linear models, for example, classically rely on constant treatment effect assumptions, or treatment effects defined by interaction terms. Even when they are relaxed, they do not typically directly model *idiosyncratic variation*, i.e. variation beyond that expressed by the model. We propose a different approach that extends the Neymanian randomization framework to explicitly allow both for treatment effect variation explained by covariates, known as *systematic variation*, and for this unexplained idiosyncratic treatment effect variation. Our perspective enables estimation and testing of impact variation under very weak modeling assumptions without loss of substantial power.

Our approach leads to two practical results beyond that of providing intuitive tests and estimators for these different types of variation. First, we combine estimates of systematic impact variation with sharp bounds on overall treatment variation to obtain bounds on the proportion of total impact variation explained by a given model—this is essentially an R^2 for treatment effect variation. We believe this measure could be quite useful in understanding and describing treatment impacts. Second, as a side-benefit, by using covariates to partially account for the correlation of potential outcomes problem, we exploit our perspective to sharpen the bounds on the finite-sample variance of the average treatment effect estimate itself. As long as the treatment effect varies across observed covariates, the resulting bounds are sharper than the current sharp bounds in the literature. We apply these ideas to a large randomized evaluation of a job training program as well as the Head Start Impact Study, showing that these results are meaningful in practice.

Purpose / Objective / Research Question / Focus of Study:

Description of the focus of the research.

The goal of this work is to create a framework that (1) provides applied researchers with a set of practical tools and (2) clearly lays out all the relevant assumptions for assessing treatment effect variation. We build this from the ground up, laying out a randomization-based framework for characterizing and understanding treatment effect heterogeneity in a range of settings, including observational studies. Following a long tradition in statistics, we use potential outcomes (Rubin, 1974; Neyman, 1923 [1990]) as the building blocks of this framework, allowing us to clearly separate the quantities of interest from the estimation methods. We decompose overall treatment effect heterogeneity into two components, the *systematic component*, impact variation explained by covariates, and the *idiosyncratic component*, impact variation not explained by covariates.

This approach has several key features. First, it allows for a model-free exploration of systematic variation. Using this approach, researchers can, with simple and relatively transparent analyses, estimate and describe trends in treatment effect variation. These arguments and related inferences tend to be quite robust as they are justified by the randomization itself. Furthermore, the randomization-based methods that we proposed are highly flexible and easy to implement. A primary goal is to provide methods to make these investigations more accessible to applied researchers.

Setting: N/A.

Population / Participants / Subjects: N/A

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

We use two major datasets as test cases for the different methods we explore. First, we analyze data from the Job Search Intervention Study (JOBS II), a randomized evaluation of an intervention for unemployed workers consisting of a series of training sessions. For example analyses, see Little & Yau (1998), Jo (2002), or Jo & Stuart (2009).

Second, we analyze data from the Head Start Impact Study, a large-scale randomized evaluation of the Head Start program in which children randomized to treatment were offered a seat in a classroom in a Head Start program in fall 2002 for the 2002-2003 school year (Michael Puma et al., 2010). This study involved 4,440 children in 351 centers that were randomized to treatment or control.

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

Existing methods range from regressions with interactions (Crump, Hotz, Imbens, & Mitnik, 2008) to hierarchical models on the marginal variances under treatment and control (Bryk & Raudenbush, 1987). Such methods explore important aspects of systematic treatment effect variation, but often rest on hidden or, at the very least, opaque assumptions on the potential outcomes themselves. Our framework allows us to clearly lay out all the assumptions—both implicit and explicit—in these approaches. In so doing, we extend a long statistical literature using potential outcomes to understand the use of regression and other methods for estimating overall effects (see, e.g., (Freedman, 2008a; 2008b; Lin, 2013)). Other examples of this approach have been successful in related contexts. Schochet (2013), for example, uses the potential outcome framework to clarify regression adjustment in cluster-randomized trials. Aronow, Green, and Lee (2013) use similar randomization arguments to motivate new estimators for cluster randomized trials. We argue that, in a spirit similar to these works, by focusing on potential outcomes and the randomization itself, we can reconcile many approaches for treatment effect heterogeneity in a unified framework for data analysis, as well as provide some new approaches.

Also, *describing* treatment effect variation is a difficult problem. There are some measures, such as using site-level variation (e.g., Bloom, Porter, Weiss, & Raudenbush, (2013)) in a hierarchical models. We hope to augment these works by providing an overall measure of treatment effect variation explained by the entire model.

Statistical, Measurement, or Econometric Model:
Description of the proposed new methods or novel applications of existing methods.

We model heterogeneity directly, dividing questions about treatment effect variation into two broad categories (Djebbari & Smith, 2008): (1) variation in treatment effects across observed characteristics (systematic variation); and (2) variation in treatment effects not explained by observed characteristics (idiosyncratic variation). We use the potential outcomes notation (Rubin, 1974; Neyman, 1923 [1990]). In this setting, we observe N individuals, N_1 of whom randomly receive some encouragement to take up an active intervention (i.e., JOBS II or center-based child care) denoted by $Z_i = 1$, and N_0 of whom are do not receive this encouragement, denoted $Z_i = 0$.

We then model individual treatment effects. Heterogeneity is then a natural function of these individual effects. In particular we model potential outcomes as:

$$Y_i(1) = Y_i(0) + \tau(X_i) + \delta_i$$

with $\tau(X)$ being a treatment effect function. Here, any variability of $Y_i(1)$ given $Y_i(0)$ and X_i is due to idiosyncratic variation; the variability of $\tau(X_i)$ is systematic variation. Under this model, we evaluate heterogeneity as a two-step process: (1) model systematic variation as best one can, and (2) assess whether the model is adequate for capturing substantively meaningful trends in treatment effect. These steps rely on two different inferential procedures, both justified by the randomization itself. In fact the latter test is guaranteed valid, as it is built on permutation-style inference procedures. The former is asymptotic in a similar form to a *t*-test, as motivated by Neyman's 1923 work.

In particular, using the above, we can fit a "sharp" null model for *systematic treatment effect variation* as a linear function of the covariates:

$$\tau_i = Y_i(1) - Y_i(0) = \boldsymbol{\beta}_0^\top \boldsymbol{X}_i, \text{ for some } \boldsymbol{\beta}_0, \text{ for all } i = 1, \cdots, N.$$

Note that, while this will be suitable for most practical applications, more general forms, such as polynomials and splines, are possible. This model differs from classic regression; in our framework, the control outcome surfaces are absent and could in theory be arbitrary curves. Our model is therefore a semi-parametric relaxation of OLS.

This model is sharp. There is no error term; given this model, and given either a treatment or control potential outcome, the unobserved outcome could be imputed exactly. Any difference between an actual treatment effect and this model would therefore be idiosyncratic variation, which we can test for separately.

Usefulness / Applicability of Method:
Demonstration of the usefulness of the proposed methods using hypothetical or real data.

Generally, we believe this framework allows researchers to naturally formulate and investigate questions about heterogeneity. As part of this work we discuss practical trade-offs such as assessing potential loss of power from using this approach relative to conventional ones. We generally find that these costs are minimal.

Overall, this framework can be used to tackle three basic types of research question posed to heterogeneity: (1) how to model and test systematic variation, (2) how to test for idiosyncratic variation, and (3) how to measure the extent of variation present. All three of these questions are receiving increased attention in the education world. See, for example, the recent overview papers of Schochet, Puma, & Deke (2014) and Weiss, Bloom, & Brock (2014).

Research Design: N/A

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

We plan on using data sets that already exist, and that we already have extensive experience using.

Findings / Results:

To illustrate our methods, we examine a sequence of models explaining treatment effect variation in the Head Start Impact Study. Results are on Table 1. (Insert Table 1 here.) The first model tests for constant effects, and we reject that model: under very weak assumptions we find variation in impact. Our methods readily extend to allow for using covariates to increase precision, and this is done in the second model where our *p*-value decreases markedly. The significance of our finding of heterogeneity is substantially increased. The third and fourth models explain variation using covariates, i.e., we allow for treatment effect to vary systematically by age, DLL (Dual Language Learner) status and baseline academic skill, and we still find that there is "heterogeneity on the table" beyond that which we are modeling. This demonstrates that we do not fully understand the pattern of impacts of treatment effects. Next steps are to assess the proportion of variation explained; perhaps the remaining heterogeneity is ignorable, or a substantial fraction of the variation found. These are the sorts of analyses that our methods allow, and the sorts of questions that allow for more nuanced understanding of an overall impact.

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

Our overall framework provides a toolkit of easy-to-use tools rooted in random assignment for modeling and assessing treatment effect heterogeneity. We show that these tools are practical, reasonably powerful, and interpretable. Furthermore, due to reliance on weak assumptions, they can be integral for building solid statistical arguments for skeptical audiences.

Appendices

Not included in page count.

Appendix A. References

References are to be in APA version 6 format.

Aronow, P. M., Green, D. P., Lee, D. K. (2013). Sharp bounds on the variance in randomized experiments. *Multiple Values Selected*, *42*(3), 850–871.

Bloom, H. S., Porter, K. E., Weiss, M., & Raudenbush, S. W. (2013). Estimating Cross-site Impact Variation in the Presence of Heteroscedasticity (Slides).

Bryk, A. S., & Raudenbush, S. W. (1987). Application of hierarchical linear models to assessing change. *Psychological Bulletin*, *101*(1), 147.

Crump, R. K., Hotz, V. J., Imbens, G. W., & Mitnik, O. A. (2008). Nonparametric Tests for Treatment Effect Heterogeneity. *Review of Economics and Statistics*, *90*(3), 389–405. doi:10.1037/h0037350

Djebbari, H., & Smith, J. (2008). Heterogeneous impacts in PROGRESA. *Journal of Econometrics*, *145*(1), 64–80.

Freedman, D. A. (2008a). On regression adjustments in experiments with several treatments. *The Annals of Applied Statistics*, *2*(1), 176–196. doi:10.1214/07-AOAS143

Freedman, D. A. (2008b). On regression adjustments to experimental data. *Advances in Applied Mathematics*, *40*(2), 180–193.

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*(23), 2857–2875.

Lin, W. (2013). Agnostic notes on regression adjustments to experimental data: Reexamining Freedman’s critique. *The Annals of Applied Statistics*, *7*(1), 295–318.

Little, R. J., & Yau, L. H. (1998). Statistical techniques for analyzing data from prevention trials: treatment of no-shows using Rubin's causal model. *Psychological Methods*, *3*(2), 147.

Puma, Michael, Bell, S., Cook, R., Heid, C., Shapiro, G., Broene, P., et al. (2010). Head Start Impact Study. Final Report. *Administration for Children \& Families*.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, *66*(5), 688–701. Retrieved from http://psycnet.apa.org/journals/edu/66/5/688/

Schochet, P. Z. (2013). Estimators for Clustered Education RCTs Using the Neyman Model for Causal Inference. *Journal of Educational and Behavioral Statistics*, *38*(3), 219–238. doi:10.3102/1076998611432176

Schochet, P. Z., Puma, M., & Deke, J. (2014). *Understanding Variation in Treatment Effects in Education Impact Evaluations: An Overview of Quantitative Methods*. U.S. Department of Education.

Splawa-Neyman, J., Dabrowska, D. M., & Speed, T. P. (1990). On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9. *Statistical Science*, *5*(4), 465–472.

Weiss, M., Bloom, H. S., & Brock, T. (2014). *A Conceptual Framework for Studying the Sources of Variation in Program Effects*. MDRC.

Appendix A. Tables

Table 1. FRT p-values for the Head Start Impact Study, based on 1,000 repetitions. Models (1) and (2) correspond to a null hypothesis of constant treatment effect. Models (3) and (4) allow the treatment effect to vary across given covariates.

	(1)	(2)	(3)	(4)
p-value:	**0.027**	**0.007**	**0.005**	**0.007**
Treatment effect varies by:	—	—	age	age DLL status acad. skills
Control for covariates:	—	✓	✓	✓