Hierarchical Models for Causal Effects[1]

Avi Feller

Harvard University Department of Statistics

and

Andrew Gelman

Columbia University Department of Statistics and Department of Political Science

Abstract

Hierarchical models play three important roles in modeling causal effects: (1) accounting for data collection, such as in stratified and split-plot experimental designs; (2) adjusting for unmeasured covariates, such as in panel studies; and (3) capturing treatment effect variation, such as in subgroup analyses. Across all three areas, hierarchical models, especially Bayesian hierarchical modeling, offer substantial benefits over classical, non-hierarchical approaches. After discussing each of these topics, we explore some recent developments in the use of hierarchical models for causal inference and conclude with some thoughts on new directions for this research area.

Keywords

Hierarchical models; causal inference; Bayesian inference; interactions; varying treatment effects

Disciplines

Statistics

[1] For *Emerging Trends in the Social and Behavioral Sciences*, ed. Robert Scott and Stephen Kosslyn. We thank Jennifer Hill and Shira Mitchell for helpful comments and the National Science Foundation and the Institute of Education Sciences for partial support of this work.

1. Background on hierarchical modeling and causal inference

Historically, social scientists have estimated causal effects via classical, linear regression. When the data come from a simple randomized experiment, this can be a very sensible approach. However, when the data deviate from this ideal even slightly, continuing to use this approach can be problematic. For example, linear regression will yield misleading results if the researcher collects test scores at the student level but randomizes assigns the intervention at the classroom level. The goal of this approach is to use statistical methods that properly account for non-standard data structures and other features of the data collection process. Moreover, these methods can be extended to data not obtained from randomized experiments.

The statistical term *hierarchical modeling* has two, related, meanings. First, it can refer to modeling of hierarchical data structures: for example, students within schools or, for a non-nested example, panel or time-series cross-sectional data, in which observations can be indexed by item or by time. Second, it can refer to modeling of the parameters within a model. We will use the two formulations interchangeably; indeed, models of hierarchical data structures can generally be viewed as add-ons to non-multilevel models, by starting with a regression model with group indicators and then assigning a second-level model to the coefficients for the group indicators.

Causal inference can be formulated statistically as a missing-data or prediction problem, with the effect of a specified treatment on a specified item being the difference between the predicted outcome conditional on the treatment. In the standard notation (Neyman, 1923, Rubin, 1974), for item i there is a treatment Z_i that can equal 0 or 1, a set of pre-treatment predictors X_i, and potential outcomes Y_{i0} and Y_{i1} corresponding to what would be observed under one treatment

or the other. The causal effect of $Z_i = 1$ compared to $Z_i = 0$ is then conventionally defined as $Y_{i1} - Y_{i0}$. Unless otherwise stated, we assume the stable unit treatment value assumption and an ignorable assignment mechanism throughout.

In the present article we consider three ways in which hierarchical modeling can aid in causal inference:

1. **Accounting for data collection:** In any data analysis, it is appropriate to account for any individual or group characteristics that are predictive of treatment assignment and inclusion in the dataset. When there are many such variables (or if these design variables include categorical predictors with many levels), multilevel modeling is a stable way to model and adjust for these, in the same way that it is appropriate to condition on all variables that affect the probability of inclusion in a sample survey (e.g., Gelman, 2007).
2. **Adjusting for unmeasured covariates:** In observational studies it is typically necessary to adjust for differences between treated and control items. If the observations are structured (for example, with longitudinal, panel, or time-series cross-sectional data), multilevel modeling can yield more efficient estimates than classical no-pooling estimates.
3. **Modeling variation in treatment effects:** Often there is interest not just in an average treatment effect but also in how the effect varies across the population. In the above notation, we can model the expected treatment effect as a function of pre-treatment covariates x, and we can also model the unexplained variance in the treatment effect.

We devote most of our space to the third of these issues because we feel it is otherwise under-emphasized both in formal statistics and in applications. In standard presentations of causal inference there is a quick jump from treatment effects defined very generally to them being

assumed constant or estimated only as averages, and we believe there is the potential to learn much more from data.

There is a long history of the use of hierarchical models for estimating causal effects, especially in education statistics (Bryk and Raudenbush, 2002). One reason for a revival of this topic now is that statisticians are increasingly able and willing to fit complex regression models using regularization to handle large numbers of predictors and arbitrary nonparametric curves (e.g., Tibshirani, 1996; Rasmussen and Williams, 2006). We as a field are moving away from inferences for single parameters in linear models and toward an acceptance of complex structures that can only be estimated with large uncertainties.

2. Hierarchical Modeling to Account for Data Collection

There is a general principle in statistics that the information used in the design of data collection—the variables that are predictive of an item being included in the study, and the variables that are predictive of treatment assignment—should be included in the analysis. This principle is supposed to be followed in classical design-based analyses as well as in model-based or Bayesian analyses (Gelman et al., 2003).

Design information can be included in various ways, including survey weighting, regression modeling, and poststratification. Moreover, specific tools can be used and interpreted using different statistical paradigms. For example, propensity scores (a form of model-based estimate of the probabilities of treatment assignment; Rosenbaum and Rubin, 1983) can be used to construct weights to estimate average treatment effects in randomized experiments in which the probability of assignment to treatment varies. Alternatively, the analysis of data from a randomized block design can incorporate block indicators. In either case, the information that goes into the design is being used in some way in the analysis.

This section discusses the use of hierarchical models to incorporate information from several types of experimental designs: stratified designs, cluster-randomized designs, split-plot designs, and longitudinal designs. For a more complete discussion of hierarchical models to account for data collection, see Hill (2013).

Completely Randomized Experiment

In a completely randomized experiment, the simplest estimate of the average treatment effect is via non-multilevel regression:

$$y_i \sim N(\alpha + \beta x_i + \tau z_i,\ \sigma_y^2)$$

where $\hat{\tau}$ is an estimate of the treatment effect under certain assumptions.[2] The coefficients in this model can be allowed to vary by group, whether or not the grouping is part of the design of the study.

Stratified Experiments

In a stratified or blocked experiment, random assignment depends on one or more observed covariates; for example, treatment is randomly assigned to half of the men and half of the women in a study population. Failing to account for this stratification in subsequent analysis leads to bias, since, in Rubin's terminology, treatment assignment is not unconditionally ignorable. That is, the outcomes and treatment assignment are not marginally independent. We can account for this in a multilevel setting by allowing the intercept, α, to vary by group:

[2] For a discussion of issues surrounding regression in the context of analyzing randomized experiments, see Imbens and Rubin (2014).

$$y_i \sim N(\alpha_{g[i]} + \beta x_i + \tau z_i,\ \sigma_y^2)$$

$$\alpha_g \sim N(\mu_\alpha, \sigma_\alpha^2)$$

where x_i are individual-level covariates; z_i indicates treatment assignment; and $\alpha_{g[i]}$ is the group-level intercept corresponding to individual i's group, g. In practice, this approach differs most starkly from classical non-hierarchical regression when the number of groups is large, though this approach is still sensible for only a few groups.

As $\sigma_\alpha \to \infty$, the model reduces to the classical "fixed effects" or *no pooling* estimate, with separate intercept estimates for each group. As $\sigma_\alpha \to 0$, the model reduces to *complete pooling* and fixes all α_g at a common α. Our preferred *partial pooling* approach is a compromise between these two estimates, partially pooling the group level parameters, α_g toward the group-level mean, μ_α, by an amount that depends on σ_α and the sample size of each group.

The variation of the intercepts α_g is dictated by the experimental design, but in general the coefficients β and τ can be allowed to vary as well. As we discuss later in this chapter, variation in the treatment effect τ can be of substantive interest. The resulting hierarchical model includes a covariance matrix for the distribution of (α, τ) or (α, β, τ) that must be estimated from data. In addition, such a model should include group-level predictors where appropriate to model predictable variation among the groups beyond what is explained by the individual-level predictors.

A final technical note is that these models assume that the probability of treatment assignment is constant across groups, that is $\bar{z}_g = \bar{z}$ for all g. If this does not hold, including $\bar{z}_g$ as a group-level predictor is often suitable to correct for the relevant modeling issues. See Bryk and Raudenbush (2002), Gelman and Hill (2007), and Hill (2013) for discussions.

Cluster-Randomized Experiments

In stratified experiments, randomization depends on the group but is still applied at the individual level. In cluster-randomized experiments, every unit in the cluster receives the same treatment; in other words, randomization occurs at the group level.[3] This design is common in the social sciences. Examples include public health interventions rolled out by city (Imai, King, and Nall, 2009, Hill and Scott, 2009), political advertising applied at the media market level (Green and Vavreck, 2007), and educational interventions at the classroom or whole-school level (Raudenbush, Martinez, and Spybrook, 2007).

Extending hierarchical models to such experiments is simple—we use the same model as for a stratified experiment but include treatment assignment as a group-level rather than individual-level predictor:[4]

$$y_i \sim N(\alpha_{g[i]} + \beta x_i,\ \sigma_y^2)$$

$$\alpha_g \sim N(\mu_\alpha + \tau z_g, \sigma_\alpha^2)$$

where z_g is the group-level treatment indicator. Moreover, adding group-level covariates beyond treatment assignment is straightforward.[5]

Split-Plot Design

[3] Some texts have slightly different definitions of cluster-randomized experiments, instead treating the term as an umbrella for all randomizations that depend on the group. See Hill (2013).
[4] In this simple case, the model is algebraically equivalent to the hierarchical model in the previous section, with the treatment effect at the individual level, and the group-level random effect. Nonetheless, we find this formulation useful for motivating more complex settings.
[5] See Imai et al. (2009) and Hill and Scott (2009) for discussion of design- versus model-based inference for cluster-randomized designs.

One increasingly common extension of cluster-randomized experiments is a split-plot design, in which different treatments are applied at different levels. For example, Sinclair, McConnell, and Green (2012) conducted a randomized voter mobilization experiment in which everyone in the treated zip code received a voter mobilization mailing, but only some households received a follow-up phone call, a natural multilevel setting. Another example is Hong and Raudenbush (2006), who assessed the effect of holding back low-achieving kindergarten students on classrooms and schools—a setting of students within classrooms within schools.

In these examples, the "group-level treatment" is simply the count ofhow many individuals in that group receive the treatment (in this case, whether 10 percent or 20 percent of the class is held back). The setup applies more generally, however: treatments at different levels can be completely different. Extending the above models yields,

$$y_i \sim N(\alpha_{g[i]} + \beta x_i + \tau_1 z_{1,i},\ \sigma_y^2)$$

$$\alpha_g \sim N(\mu_\alpha + \tau_2 z_{2,g}, \sigma_\alpha^2)$$

where $z_{1,i}$ indicates treatment at the first level (that is, individuals), and $z_{2,g}$ indicates treatment at the second level (groups).

Longitudinal and Repeated Measurements

Longitudinal designs often mimic cluster-randomized experiments: there are multiple observations for the sample individual, who is either assigned to treatment or control. In practice, longitudinal data analysis requires consideration of complex correlation and missing-data issues (see Diggle,et al., 2002, and Van der Laan and Robins, 2003). For the sake of illustration,

however, we note the connection between the previous models and a simple hierarchical model that includes information about collecting observations within waves:

$$y_i \sim N(\alpha_{j[i]}^{indiv} + \alpha_{k[i]}^{time},\ \sigma_y^2)$$

$$\alpha_j^{indiv} \sim N\left(\mu_{indiv} + \beta x_j + \tau z_j,\ \sigma_{indiv}^2\right)$$

$$\alpha_k^{time} \sim N(\mu_{time},\ \sigma_{time}^2)$$

Extensions of this setup allow for regimes that vary over time. Examples of multilevel modeling in this setting include Zajonc (2012), who investigates student tracking, and Hong and Raudenbush (2007), who analyze the effects of time-varying instructional treatments.

3. Hierarchical Modeling to Adjust for Unmeasured Covariates

In observational studies, researchers must take extra steps when comparing treated and control units. In general, matching and propensity score-based approaches can help to restrict the data to a regime under which regression-based approaches can be used to control for systematic differences between treatment and control units (see Rubin and Imbens, 2014). Just as with more complex experimental designs, propensity scores models can often be improved by a hierarchical structure (Arpino and Mealli, 2011).

The econometrics literature has focused on an alternative approach for observational studies, especially when estimating causal effects for panel data. In the simplest case, observations are assumed to come from a simple linear model, in which each individual has a single, unobserved trait, such as ability, that is assumed to be constant over time (see Wooldridge, 2010, for a discussion of the relevant assumptions). Consider the example of

estimating the effect of union membership on (log) wages using individual-level panel data (e.g., Angrist and Pischke, 2009):

$$y_i = \alpha_{j[i]}^{indiv} + \alpha_{t[i]}^{time} + \tau z_i + \varepsilon_i$$

where z_i denotes union membership and $\alpha_{j[i]}^{indiv}$and $\alpha_{t[i]}^{time}$ are intercept estimates for individual j and time t respectively. As before, the key inferential question is the choice of the appropriate model for α_i, including whether the relevant assumptions are applicable. There is a long literature in econometrics comparing the no-pooling and partial-pooling estimates in this case, which correspond to the so-called "panel fixed effects" and "panel random effects" estimators. As with randomized experiments, the partially pooled estimates are often more efficient than the no pooling estimates (Hausman, 1978, Wooldridge, 2010).

4. Hierarchical Modeling to Account for Treatment Effect Variation

The above models all center on the estimation of constant additive treatment effects; that is: $Y_{1i} - Y_{0i} = \tau_i = \tau$, for units $i = 1, \ldots, n$. In practice, however, we know that everything varies, especially in the social sciences, making this an overly restrictive modeling assumption.

The key point is not just that treatment effects vary, but that we can both predict this variation and use it to better understand the intervention of interest (Ding, Feller, and Miratrix, 2014). For example, practitioners implementing a program face budget constraints and must target resources to a subset of the population (see, for example, Dehejia, 2005, Imai and Strauss, 2011). Or policymakers are interested in the effects of a given intervention on the distribution of resources in society (Bitler, Gelbach, and Hoynes, 2003). In other settings, variation in a causal effect is itself of interest (Gelman and Huang, 2008).

Even when not formally in the model, treatment effect variation is implicitly recognized. Once we start talking about average treatment effects (ATE), local average treatment effects (LATE), and the like, we are implicitly talking about interactions and varying treatment effects. Were the treatment effect truly constant across all units, we could just speak of "the treatment effect" without having to specify which cases we are averaging over. Any discussion of particular average treatment effects is relevant because treatment effects vary; that is, the treatment effect interacts with pretreatment variables.

Before turning to estimation, it is important to note that there are some fundamental limitations to inference about varying treatment effects. First, as with all causal inference problems, we can never observe the joint distribution of Y_0 and Y_1, only their marginal distributions. Therefore, we never actually observe τ_i directly and, in general, can only make inference about its full distribution via additional modeling assumptions, though we can almost always identify the expectation of τ_i given a set of covariates (see Ding et al. (2014) for a discussion). Second, inference about *systematic variation* in τ_i (i.e. the variation that can be explicitly modeled) does not necessarily imply inference about underlying causal mechanisms, in the same way that correlation does not necessarily imply causation. See Gerber and Green (2012) for illustrative political examples.

With those caveats, we now turn to modeling treatment effects that vary. Ideally, such modeling should be grounded in theory. But, as we discuss below, even when guided by strong substantive considerations, researchers still face a broad range of non-trivial modeling choices. We therefore explore a broad range of models for treatment effect variation, beginning with the simplest—and by far most widely used—model of treatment effect variation: variation in the mean treatment effect conditional on covariates. We then turn to models of treatment effect

variation in second and higher-order moments, and to treatment effect variation on the overall outcome distribution. Finally, we discuss nonparametric methods for estimating treatment effect variation. Throughout we focus on the advantages of a Bayesian hierarchical approach over non-multilevel regression modeling.

Varying Treatment Effects on the Mean Outcome

We now return to the hierarchical model for stratified experiments from Section 2. In that setting, variation in the treatment effect, τ, was reflected in the model. We can relax that assumption in the following model:

$$y_i \sim N(\alpha_{g[i]} + \beta x_i + \tau_{g[i]} z_i,\ \sigma_y^2)$$

$$\begin{pmatrix} \alpha_g \\ \tau_g \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_\alpha \\ \mu_\tau \end{pmatrix}, \begin{pmatrix} \sigma_\alpha^2 & \rho\sigma_\alpha\sigma_\tau \\ \rho\sigma_\alpha\sigma_\tau & \sigma_\tau^2 \end{pmatrix} \right)$$

which partially pools all the group-specific treatment effect estimates toward the overall mean, μ_τ, while allowing the group-specific intercept and treatment effect to co-vary (see Bryk and Raudenbush, 2002, Gelman and Hill, 2007)

As in the previous sections, the classical approach is mathematically equivalent to the hierarchical model in which $\sigma_\alpha \to \infty$ and $\sigma_\tau \to \infty$, which implies no pooling both for the varying intercepts, α_g and the varying slopes, τ_g. This latter choice is especially problematic in the context of treatment effect interactions, implying that an interaction effect of zero is as likely as an arbitrarily large interaction—an obviously absurd statement (Dixon and Simon, 2001, Simon et al., 1997, Simon, 2002).

Moreover, this no-pooling approach can prove especially problematic in the context of trying to estimate multiple weak signals. First, consider the issue of statistical power: imagine that a researcher is interested in treatment effect variation across two equally-sized groups and that the true interaction effect is half as large as the overall effect. Since there are half as many people in each subgroup, however, the precision decreases by a factor of four relative to the precision for the main effect (see Simon, 2007). If the study is powered so that the main effect is barely statistically significant, as is typical in social science applications, then detecting the interaction effect is substantially less likely. A practical consequence is that applied researchers often look for interactions across many covariates. In a non-hierarchical setting, this creates a classic multiple testing problem, as well as a strong incentive for "specification searches" (Fink, McConnell, and Vollmer, 2011, Pocock et al., 2002). Pre-analysis plans that specify which subgroups will be analyzed before running the experiment mitigate this issue, but do not completely resolve the multiple comparisons problem in the no pooling model.

A Bayesian hierarchical approach is not a panacea for weak signals and multiple comparisons, but it does avoid some of the worst pitfalls that these create. In short, multiple comparisons can be thought of as multiple analyses in parallel, a natural opportunity for hierarchical modeling or meta-analysis. The resulting shrinkage from hierarchical Bayesian inference automatically reduces the "false positives" problem inherent in multiple classical inferences. While this is true, pre-specifying interaction effects is still important in a Bayesian setting: if a researcher believes that an interaction effect is important enough to be pre-specified, the researcher's prior distribution for that effect is more diffuse (i.e., there is more mass far from zero) than for an interaction effect that is chosen "post-hoc" (Simon, 2009).

To date, we are aware of very few analyses in the social sciences that partially pool treatment effect estimates across groups. One promising exception is in the analysis of so-called multi-site trials, which are common in social policy and education evaluations (see, for example, Bloom, Raudenbush, and Weiss, 2013). Another is recent work by Imai and Ratkovic (2013), who use Lasso—rather than a Bayesian hierarchical model—to regularize these treatment effect estimates.

Variance Components Models of Treatment Effects

As in the previous section, varying treatment effects can be modeled as interactions. But such models can also be rewritten in terms of variance components (Bryk and Raudenbush, 2002; Gelman and Hill, 2007):

$$y_i = \alpha + \beta x_i + \tau z_i + e_i$$

$$e_i = \nu_{g[i]} + \xi_{g[i]} z_i + \varepsilon_i$$

where $\nu_{g[i]} \sim N(0, \sigma_\nu^2)$; $\xi_{g[i]} \sim N\left(0, \sigma_\xi^2\right)$; and $\varepsilon_i \sim N(0, \sigma_\varepsilon^2)$. If $\nu_{g[i]}$ and $\xi_{g[i]}$ were modeled as a bivariate Normal, this would be exactly equivalent to the previous model.

Written in this way, it is clear that a much richer class of models are possible for the error term, e_i. The first such models were proposed in a seminal paper by Bryk and Raudenbush (1988), who model the error term, e_i, as a function of continuous (standardized) covariates:[6]

$$e_i = \xi_i x_i + \varepsilon_i$$

Therefore, positive values of x_i are associated with larger variance; for example, high-ability students will show more variability in test scores than low-ability students.

[6] The authors extend this to multiple covariates. There is a large literature on comparing variances in experimental settings. See, for example, Cox (1958) and Cox (1998).

Applying this model directly to the case of a binary treatment indicator implies that the treatment group will have larger variance than the control group. Gelman (2004) calls this an *additive treatment error* model, which would naturally arise in a situation with a new, active intervention that has a variable impact. However, the opposite situation—*subtractive treatment error*—is also possible; for example, in before-after data, if an active intervention is applied in the "before" period and the treatment consists of dropping this intervention in the "after" period (i.e. for incumbency; see Gelman and Huang, 2008).[7] Of course, a richer set of variance components models are possible, such as allowing the variance to depend on a complex function of pre-treatment covariates. For example, in the context of a multilevel experimental setting, Kim and Seltzer (2011) fit hierarchical regressions on the variance components themselves.

Such models can prove difficult to fit in practice. As Bryk and Raudenbush (1988) observe, fat tails and other departures from the assumed parametric model can create significant complications. We believe that, in these cases, it is especially important to check that the model is a good fit to the observed data (Gelman et al., 2013; Ch. 6).

Variation in the Effect of Treatment on Quantiles

Thus far, our models have focused on the first and second moments of the outcome distributions. An alternative modeling strategy, especially popular in economics, is instead to compare percentiles of the marginal distributions, Q_{y0} and Q_{y1} (e.g., Angrist and Pischke, 2008). For example, Bitler et al. (2003) investigate the effects of welfare reform on the entire wage distribution, not just the average for a subset of the population. Similarly, Dominici, Zeger,

[7] Gelman (2004) also discusses a "replacement treatment error" model, in which the treatment replaces a random error component from the "before" period.

Parmigiani, Katz, and Christian (2006) estimate effect of nutrients on low-birth-weight infants, rather than on average weights.

Modeling quantiles is often more challenging than modeling means. A growing literature has focused on this estimation challenge in a Bayesian context (Chamberlain and Imbens, 2003, Lancaster and Jun, 2009, Taddy and Kottas, 2010). Moreover, many of the hierarchical modeling approaches described above can be extended to quantile regression (Reich, Bondell, and Wang, 2010), simply replacing the mean by the relevant quantile.

Advances in Models for Treatment Effect Variation

Flexible Parametric Methods for Treatment Effect Variation

As in other areas of Bayesian modeling, new developments in models for treatment effect variation across groups focus on richer parameterizations of the simpler models discussed above. For example, there is a growing literature on specifying the prior variance for interaction effects, (see Sargent and Hodges, 1997, and Hodges et al., 2007). One promising approach is the use of nonparametric prior distributions (Sivaganesan, Laud, and Müller, 2010) or in highly flexible hierarchical array priors (Volfovsky and Hoff, 2012) for these interaction terms.

Another extension is to allow for treatment effects to vary across continuous covariates. Researchers rarely—if ever—have a substantive reason to make a particular assumption of the parametric form of the interaction. For example, we might have a strong substantive reason why a given treatment will become increasingly effective for higher income voters. However, theory is unlikely to tell us that this relationship is linear, quadratic, or exponential—there is no strong reason ahead of time to believe that treatment effect is more likely to increase with income rather than log-income (Beck and Jackman, 1998). Some researchers seek to avoid this problem by

discretizing their continuous variable, but this simply pushes the problem back to a specification search of a different kind, in which researchers find cutpoints that lead to the best results (Assmann, et al., 2000). Flexible models for continuous covariates, such as splines and Gaussian processes, offer a promising solution to this issue (Feller and Holmes, 2009).

Finally, we would also like to see a class of models in which treatments with larger main effects naturally have larger variation. As Cox (1984) observed, "large component main effects are more likely to lead to appreciable interactions than small components. Also, the interactions corresponding to larger main effects may be in some sense of more practical importance." See also Gelman, 2004). Bien, Taylor, and Tibshirani (2012) implement a model that respects this hierarchy restriction in the context of the Lasso. To our knowledge, however, there are no such models in a Bayesian setting.

Nonparametric Response Surface Modeling

An alternative approach is based on directly modeling the distributions of Y_0 and Y_1, also known as *response surface modeling*. For example, Hill (2011) used Bayesian additive regression trees (BART) to model treatment effect interactions (see also Green and Kern, 2012, and Imai and Strauss, 2011). More broadly, cutting-edge Bayesian nonparametric methods, such as Gaussian processes, can be used to flexibly model the response surface (e.g., Tokdar, 2013).

Variation across Latent Subgroups

Finally, there is an increasing focus on modeling treatment effect variation across latent or partially-observed subgroups. This is especially promising for *principal strata*, subgroups defined by the joint distribution of intermediate outcomes, such as treatment take-up, under

treatment and control (Frangakis and Rubin, 2002). Since many researchers already fit such models in a Bayesian framework (Hirano et al., 2000; Imbens and Rubin, 2014), it is straightforward to extend these models to a multi-level setting.

5. Conclusions

When doing causal inference for experiments and observational studies, the ubiquitous statistical challenge is to control for systematic pre-treatment differences between treatment and control groups. Multilevel modeling arises here for three reasons: in fitting the statistical model to multilevel data structures, as a tool for regularizing in matching or regression models with large numbers of predictors, and for modeling variation in treatment effects.

Bibliography

Arpino, B., and Mealli, F. (2011). The specification of the propensity score in multilevel observational studies. *Computational Statistics and Data Analysis*, 55(4), 1770-1780.

Assmann, S. F., Pocock, S. J., Enos, L. E., and Kasten, L. E. (2000). Subgroup analysis and other (mis) uses of baseline data in clinical trials. *Lancet*, *355*(9209), 1064–1069.

Beck, N., and Jackman, S. (1998). Beyond linearity by default: Generalized additive models. *American Journal of Political Science*, 596–627.

Bertrand, M., Duflo, E., and Mullainathan, S. (2004). How much should we trust differences-in-differences estimates? *The Quarterly Journal of Economics*, *119*(1), 249–275.

Bien, J., Taylor, J., and Tibshirani, R. (2012). A lasso for hierarchical interactions. *arXiv preprint arXiv:1205.5050*.

Bitler, M., Gelbach, J., and Hoynes, H. (2003). What mean impacts miss: Distributional effects of welfare reform experiments. *American Economic Review*, *96*(4), 988–1012.

Bloom, H. S., Raudenbush, S. W., and Weiss, M. (2013) Estimating variation in program impacts: theory, practice and applications. *MDRC Working Paper.*

Bryk, A. S., and Raudenbush, S. W. (1988). Heterogeneity of variance in experimental studies: A challenge to conventional interpretations. *Psychological Bulletin*, *104*(3), 396–404.

Bryk, A. S., and Raudensbush, S. W. (2002). *Hierarchical Linear Models: applications and data analysis methods* (second edition). Thousand Oaks, CA: Sage Publications.

Chamberlain, G., and Imbens, G. W. (2003). Nonparametric applications of Bayesian inference. *Journal of Business and Economic Statistics*, *21*(1), 12–18. doi:10.1198/073500102288618711

Cox, D. R. (1984). Interaction. *International Statistical Review*, *52*(1), 1–31. doi:10.2307/1403235

Dehejia, R. H. (2005). Program evaluation as a decision problem. *Journal of Econometrics*, *125*(1-2), 141–173. doi:10.1016/j.jeconom.2004.04.006

Diggle, P., Heagerty, P., Liang, K. Y., and Zeger, S. (2002). Analysis of longitudinal data. Oxford University Press.

Ding, P., Feller, A., and Miratrix, L. (2014) Randomization inference for treatment effect variation. Unpublished manuscript.

Dixon, D. O., and Simon, R. (1991). Bayesian subset analysis. *Biometrics*, 871–881.

Dominici, F., Zeger, S. L., Parmigiani, G., Katz, J., and Christian, P. (2006). Estimating percentile-specific treatment effects in counterfactual models: a case-study of micronutrient supplementation, birth weight and infant mortality. *Journal of the Royal Statistical Society. Series C. Applied Statistics*, *55*(2), 261–280. doi:10.1111/j.1467-9876.2006.00533.x

Fink, G., McConnell, M., and Vollmer, S. (2011). Testing for heterogeneous treatment effects in experimental data: False discovery risks and correction procedures.

Frangakis, C. E. and Rubin, D. B. (2002) Principal Stratification in causal inference. *Biometrics*, *58*(1): 21-29.

Gelman, A. (2004). Treatment effects in before-after data. In *Applied Bayesian modeling and causal inference from incomplete-data perspectives* (pp. 195–202). Chichester: Wiley. doi:10.1002/0470090456.ch18

Gelman, A. (2007). Struggles with survey weighting and regression modeling (with discussion). *Statistical Science,* 22, 153-188.

Gelman, A., and Huang, Z. (2008). Estimating incumbency advantage and its variation, as an

example of a before–after study. *Journal of the American Statistical Association*, *103*(482), 437–446. doi:10.1198/016214507000000626

Gerber, A. S. and Green, D. P. (2012) Field experiments: design, analysis, and interpretation. New York: W. W. Norton and Company.

Gerber, A. S., Green, D. P., and Larimer, C. W. (2008). Social pressure and voter turnout: Evidence from a large-scale field experiment. *American Political Science Review*, *102*(01). doi:10.1017/S000305540808009X

Gerber, A. S., Green, D. P., and Larimer, C. W. (2010). An experiment testing the relative effectiveness of encouraging voter participation by inducing peelings of pride or shame. *Political Behavior*, *32*(3), 409–422. doi:10.1007/s11109-010-9110-4

Green, D. P., and Kern, H. L. (2012). Modeling heterogeneous treatment effects in survey experiments with Bayesian additive regression trees. *Public opinion quarterly*, *76*(3), 491–511.

Green, D. P., and Vavreck, L. (2007). Analysis of cluster-randomized experiments: A comparison of alternative estimation approaches. *Political Analysis*, *16*(2), 138–152. doi:10.1093/pan/mpm025

Hausman, J. A. (1978). Specification tests in econometrics. *Econometrica*, 1251-1271.

Hill, J. L. (2011). Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, *20*(1), 217–240. doi:10.1198/jcgs.2010.08162

Hill, J. L. (2013). Multilevel models and causal inference. *The SAGE Handbook of Multilevel Modeling,* Ed. by Scott, M., Simonoff, J., and Marx, B.

Hill, J., and Scott, M. (2009). Comment: The essential role of pair matching. *Statistical Science*, *24*(1), 54–58. doi:10.1214/09-STS274A

Hirano, K., Imbens, G. W., Rubin, D. B., Zhou X.-H. (2000). Assessing the effect of an influenza vaccine in an encouragement design. *Biostatistics 1*(1): 69-88.

Hodges, J. S., Cui, Y., Sargent, D. J., and Carlin, B. P. (2007). Smoothing Balanced Single-Error-Term Analysis of Variance. *Technometrics*, *49*(1), 12–25. doi:10.1198/004017006000000408

Hong, G., and Raudenbush, S. W. (2006). Evaluating kindergarten retention policy. *Journal of the American Statistical Association*, *101*(475), 901–910. doi:10.1198/016214506000000447

Hong, G., and Raudenbush, S. W. (2007). Causal inference for time-varying instructional treatments. *Journal of Educational and Behavioral Statistics*, *33*(3), 333–362. doi:10.3102/1076998607307355

Imai, K., and Ratkovic, M. (2013). Estimating treatment effect heterogeneity in randomized program evaluation. *The Annals of Applied Statistics*, *7*(1), 443–470. doi:10.1214/12-AOAS593

Imai, K., and Strauss, A. (2011). Estimation of heterogeneous treatment effects from randomized experiments, with application to the optimal planning of the get-out-the-vote campaign. *Political Analysis*, *19*(1), 1–19. doi:10.1093/pan/mpq035

Imai, K., King, G., and Nall, C. (2009). The essential role of pair matching in cluster-randomized experiments, with application to the Mexican Universal Health Insurance Evaluation. *Statistical Science*, *24*(1), 29–53. doi:10.1214/08-STS274

Imbens, G. and Rubin, D. (2014) *Causal Inference*. Draft textbook.

Kim, J., and Seltzer, M. (2011). Examining heterogeneity in residual variance to detect differential response to treatments. *Psychological Methods*, *16*(2), 192–208. doi:10.1037/a0022656

Neyman, J. (1923). On the application of probability theory to agricultural experiments. Essay on principles. Section 9. Translated and edited by D. M. Dabrowska and T. P. Speed. *Statistical Science* 5, 463-480 (1990).

Lancaster, T., and Jun, S. J. (2009). Bayesian quantile regression methods. *Journal of Applied Econometrics*, *25*(2), 287–307. doi:10.1002/jae.1069

Pocock, S. J., Assmann, S. E., Enos, L. E., and Kasten, L. E. (2002). Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: current practiceand problems. *Statistics in Medicine*, *21*(19), 2917–2930. doi:10.1002/sim.1296

Rasmussen, C. E., and Williams, C. K. I. (2006). *Gaussian Processes for Machine Learning.* Cambridge, Mass.: MIT Press.

Raudenbush, S. W., Martinez, A., and Spybrook, J. (2007). Strategies for improving precision in group-randomized experiments. *Educational Evaluation and Policy Analysis*, *29*(1), 5–29. doi:10.3102/0162373707299460

Reich, B. J., Bondell, H. D., and Wang, H. J. (2010). Flexible Bayesian quantile regression for independent and clustered data. *Biostatistics*, *11*(2), 337–352. doi:10.1093/biostatistics/kxp049

Rosenbaum, P. R., and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*(1), 41–55.

Rothwell, P. M. (2005). Treating individuals 2-subgroup analysis in randomised controlled trials: importance, indications, and interpretation. *The Lancet*, *365*(9454), 176–186.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational psychology*, *66*(5), 688–701.

Sargent, D. J., and Hodges, J. S. (1997). Smoothed ANOVA with application to subgroup analysis. *Research Report rr2005-018, Department of Biostatistics, University of Minnesota*.

Simon, R. (2002). Bayesian subset analysis: application to studying treatment-by-gender interactions. *Statistics in Medicine*, *21*(19), 2909–2916. doi:10.1002/sim.1295

Simon, R. M. (2007). Subgroup analysis. In *Wiley Encyclopedia of Clinical Trials*. John Wiley and Sons, Inc.

Sinclair, B., McConnell, M., and Green, D. P. (2012). Detecting spillover effects: Design and analysis of multilevel experiments. *American Journal of Political Science*, *56*(4), 1055–1069. doi:10.1111/j.1540-5907.2012.00592.x

Sivaganesan, S., Laud, P. W., and Müller, P. (2010). A Bayesian subgroup analysis with a zero-enriched Polya Urn scheme. *Statistics in Medicine*, *30*(4), 312–323. doi:10.1002/sim.4108

Taddy, M. A., and Kottas, A. (2010). A Bayesian nonparametric approach to inference for quantile regression. *Journal of Business and Economic Statistics*, *28*(3), 357–369. doi:10.1198/jbes.2009.07331

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society B,* 58, 267-288.

Tokdar, S. T. (2013). Causal Analysis of Observational Data with Gaussian Process Potential Outcome Models. *Presentation at the 2013 Joint Statistical Meetings.*

Van der Laan, M. J., and Robins, J. M. (2003). Unified methods for censored longitudinal data and causality. Springer.

Volfovsky, A., and Hoff, P. D. (2012, August 8). Hierarchical array priors for ANOVA decompositions. *arXiv.org*. Retrieved from http://arxiv.org/pdf/1208.1726v1.pdf

Wooldridge, J. M. (2010). *Econometric analysis of cross section and panel data (Second Edition)*. MIT press.

Zajonc, T. (2012). Bayesian inference for dynamic treatment regimes: Mobility, equity, and efficiency in student tracking. *Journal of the American Statistical Association*, *107*(497), 80–92. doi:10.1080/01621459.2011.643747