Using Hierarchical Models to Estimate Heterogeneous Effects

Joshua Alley
Assistant Professor
University College Dublin
joshua.alley@ucd.ie

February 16, 2024

Abstract

This note describes how and when to use Bayesian hierarchical models to estimate heterogeneous effects. While an ample literature suggests that hierarchical models provide helpful regularization and information about how effects vary, political scientists rarely use them to estimate heterogeneous effects. Doing so is simple, however. To start, researchers specify groups based on quantities of interest such as demographics, context, and policy relevance. Then, researchers fit a hierarchical model where treatment slopes and intercepts vary across groups. This captures systematic and random variation in heterogeneous effects, estimates effects within each group, and measures effect variance. Hierarchical modeling provides an intermediate tool between interactions or subgroup analyses and machine learning approaches to discovering complex heterogeneity. It is more flexible than interactions and reduces the risk of underpowered subgroup comparisons. At the same time, it is more theoretically informed and interpretable than some machine learning approaches, as well as easier to implement in small datasets. Researchers can thus use hierarchical models alongside other approaches to understand heterogeneous effects for scholarship and policy.

1 Introduction

Whether in observational or experimental studies, every independent variable social scientists examine impacts some units differently than others. Common estimands aggregate heterogeneous effects.[1] Average effects can be useful, but they often obscure interesting and important variation.

As a result, understanding heterogeneous effects is essential for policy and scholarship. Estimating heterogeneity allows scholars to clarify the connection between their independent variable and outcome. Policymakers can maximize the impact of finite resources with targeted interventions, for example by providing job training to individuals who are more likely to benefit.

This letter explains how and when to use hierarchical models to estimate heterogeneous effects. I identify when researchers can profitably use hierarchical models, and when other tools make more sense. A large statistics literature suggests that Bayesian hierarchical models are a useful tool for heterogeneous effects estimation (e.g., Feller and Gelman (2015); McElreath (2016); Dorie et al. (2022)). Hierarchical models are particularly good at capturing variation across groups and levels when there are multiple potential modifiers, which is becoming more plausible as social scientists examine a variety of conditional theories. Political scientists tend to rely on interactions or machine learning tools instead. For instance, of the three applied political science citations of Feller and Gelman (2015), only Marquardt (2022) models treatment effects.

Hierarchical modeling of heterogeneous effects fills a gap between interactions and machine learning.[2] Parametric interactions and subgroup analyses are ubiquitous because they are easy to interpret. These approaches are hard to interpret with more than three dimensions

[1] For instance, Abramson, Koçak and Magazinnik (2022) note that the average marginal component effect (AMCE) of conjoint experiments gives more weight to intense preferences.

[2] Blackwell and Olson (2022) describe a lasso approach to interactions that sits between machine learning and linear regressions.

and are often underpowered (Simmons, Nelson and Simonsohn, 2011). More recent work employs random forests (Green and Kern, 2012; Wager and Athey, 2018), support vector machines (Imai and Ratkovic, 2013), and ensemble methods (Grimmer, Messing and Westwood, 2017; Künzel et al., 2019; Dorie et al., 2022). These machine learning algorithms can discover complex patterns and high-dimensional variation, but can be difficult to interpret and implement, especially in smaller social science datasets.

Using a hierarchical model is more flexible than parametric interactions but easier to implement and interpret than machine learning approaches. It preserves a simple and interpretable structure, while accommodating more factors and ameliorating the downsides of subgroup analysis via partial pooling. This facilitates argument testing. Unlike machine learning, the hierarchical approach lacks the flexibility to discover high-dimensional heterogeneity, however. Hierarchical modeling therefore works well when there are more than two modifying factors and therefore many subgroups of interest, as well as less emphasis on discovery.

There are two key steps when theory and data make using hierarchical models worthwhile. First, researchers should define groups based on potential sources of heterogeneity such as other treatments, context, demographics, or policy concerns. Second, they should estimate heterogeneous effects across those groups using a hierarchical model with varying slopes and intercepts, along with covariates that predict slopes. Modeling heterogeneous effects in this way produces interpretable results, which facilitates argument testing. It also allows researchers to examine effects within groups, compare different sources of heterogeneous effects and describe how much an effect varies.

While frequentist estimation of hierarchical models is possible, Bayesian estimation is worthwhile and preferable. Bayesian estimation provides crucial information by connecting parameters through common prior distributions, thereby regularizing estimates and propagating uncertainty. Working with posterior distributions also gives researchers more flexibility to present diverse information about how and when effects vary. While computation and coding were once a barrier to wider use of Bayesian methods, fitting a wide range of hierarchical models is straightforward with the brms package in R (Bürkner, 2017).[3] Calculating substantive effects is also simple (Arel-Bundock, N.d.).

In the remainder of this note, I describe how and when to estimate hierarchical models of heterogeneous effects and demonstrate the process by analyzing a study of how military alliances shape public support for war by Tomz and Weeks (2021). The reanalysis reveals that alliances increase support for intervention most among white men who support international engagement but are otherwise skeptical of using force.

2 Hierarchical Modeling of Heterogeneous Effects

There are two steps in hierarchical models of heterogeneous effects. First, researchers must define the groups over which an independent variable's impact changes. Unique combinations of characteristics such as other treatments, context and demographics determine groups.

Researchers should create groups based on what variation is most important and interesting. Theory, policy concerns, or normative factors are all possible motivations.

Setting groups is the most important task, because it determines what heterogeneous effects a researcher estimates. Defining groups before model fitting defines what variation is most important, links heterogeneous effects to theory, and structures modeling.[4] Defining groups without careful thought risks obfuscating results and can hinder model fitting.

There are three general approaches to defining groups. First, researchers can set groups using combinations of other treatments, especially when an intervention has several dimensions but theory emphasizes one of them. The experimental design determines groups, and the model estimates heterogeneous treatment effects. If researchers want to know how different issues shape the impact of elite foreign policy cues (Guisinger and Saunders, 2017), they could

[3] I provide example code in this note and the appendix.
[4] It also facilitates pre-registration when applicable.

define groups by issues, for instance.

A second approach uses unit, demographic and contextual factors to create groups and estimate effect heterogeneity. Here, researchers examine what factors within or around units shape their response to an independent variable. For example, Alley (2021) uses alliance characteristics to examine when alliance membership increases or decreases military spending. Other use cases include estimating how different demographic groups or geographic units respond to an intervention.

Third, researchers might emphasize policy concerns. Understanding how an intervention impacts a specific population is a common problem. Researchers might want to know if a job-training program improves employment outcomes for black women in the South, for instance.

Whether researchers use other treatments, context, or policy to determine groups, the number of grouping factors depends first on theory. There are some practical constraints, however. Using too many factors can lead to model fitting and interpretation problems by creating many small groups. How many factors is too many depends on theory and the data- some datasets can support reasonably large groups for many factors. At the other extreme, using only one grouping factor will create an unidentified model, and two will give similar results to standard interactions.

After defining groups, the second step is fitting a hierarchical model of effects within groups. The first equation links the independent variable and outcome. The second equation estimates heterogeneous effects as a function of group characteristics. The second equation is essential, as group variables add substantial information and avoid simply pooling small groups towards the overall mean. Without the group predictors, the random effects of each treatment will shift towards the mean because the information about each group depends only on how many data points it contains.

This approach can address diverse problems, but for ease of exposition consider making between-unit comparisons based on an experimental treatment. Start with N units indexed by

i, some of which receive a binary treatment T. Assume that the outcome variable y is normally distributed with mean μ_i and standard deviation σ.[5] g indexes the researcher-defined groups, which include two variables *v1* and *v2*.

The outcome for each unit is then a function of group varying intercepts α_g, an optional matrix of control variables $\mathbf{X}$, and a set of group treatment effects θ_g, which are normally distributed with mean η_g and standard deviation σ_θ. The researcher divides all units into g groups based on unique combinations of heterogeneous effect predictors $\mathbf{Z}$. Each θ parameter estimates how the treatment effect varies across the values of each variable. To capture correlations between the random intercepts and varying slopes $\rho\sigma_\alpha\sigma_\theta$, these variables should have a common multivariate normal prior.

$$y_i \sim N(\mu_i, \sigma) \tag{Likelihood}$$

$$\mu_i = \alpha + \alpha_{v1} + \theta_{v1} + T + \alpha_{v2} + \theta_{v2}T + \mathbf{X}\beta \tag{Outcome Equation}$$

$$\begin{pmatrix} \alpha_{v1} \\ \theta_{v1} \end{pmatrix} \sim N \left[\begin{pmatrix} \mu^{\alpha}_{v1} \\ \mu^{\theta}_{v1} \end{pmatrix}, \begin{pmatrix} \sigma^2_\alpha & \rho\sigma^{\alpha}_{v1}\sigma^{\theta}_{v1} \\ \rho\sigma^{\alpha}_{v1}\sigma^{\theta}_{v1} & \sigma^2_\theta \end{pmatrix} \right] \tag{Correlated Intercepts and Slopes: V1}$$

$$\begin{pmatrix} \alpha_{v2} \\ \theta_{v2} \end{pmatrix} \sim N \left[\begin{pmatrix} \mu^{\alpha}_{v2} \\ \mu^{\theta}_{v2} \end{pmatrix}, \begin{pmatrix} \sigma^2_\alpha & \rho\sigma^{\alpha}_{v2}\sigma^{\theta}_{v2} \\ \rho\sigma^{\alpha}_{v2}\sigma^{\theta}_{v2} & \sigma^2_\theta \end{pmatrix} \right] \tag{Correlated Intercepts and Slopes: V2}$$

$$\theta_g = \theta_{v1}\mathit{V1} + \theta_{v2}\mathit{V2} \tag{Group Slopes}$$

$$\alpha_g = \alpha_{v1} + \alpha_{v2} \tag{Group Intercepts}$$

This approach lets slopes vary across multiple variables within a group. The net impact of the treatment in each group depends on the linear combination of slopes in that group,

[5]Researchers should use binary, categorical and other outcome likelihoods.

specifically the θ parameters and values of each variable. Similarly, the group intercepts can be calculated as the sum of the corresponding random intercepts for each grouping variable.[6]

The above model can be fit with Bayesian or frequentist methods, but Bayesian estimation offers some useful advantages. First, it is more flexible, and including prior information can facilitate model fitting and convergence. Priors also help regularize estimates by pulling extreme groups towards the overall mean. Working with posterior distributions also also provides a wealth of information about effect heterogeneity and propagates uncertainty.

In interpreting these models, researchers should leverage the full range of information from the different parameters. First, the θ posteriors give the impact of a variable within each group. All θs reflect a systematic component from the predictors in $\mathbf{Z}\lambda$ and a random variation in slopes from σ_θ. Unless the group-level predictors are weak correlates of treatment response, the systematic component will dominate. The random variation is similar to the error term in regression- it expresses how much variation is left in addition to the systematic component.

In addition to group-specific effect estimates, a hierarchical model facilitates rich description of effects across groups. It estimates how specific factors drive differences between groups via the λ parameters. Researchers can also calculate variance in the θ parameters across groups and compare the posteriors of different θs. The σ_θ parameter summarizes the random variation. Other techniques such as interactions in OLS with robust standard errors provide less information.

3 When to Use Hierarchical Models

In deciding whether to use hierarchical models, researchers must weigh it's unique advantages and disadvantages. In general, estimating heterogeneous effects in this way has three

[6]In brms for a model with no controls and two variables modifying the impact of a treatment, the model formula is y $\sim$ 1 + (1 + treat | var1) + (1 + treat | var2). If joint variation across variables 1 and 2 is also of interest, append (1 + treat | var1:var2).

advantages. First, researchers can make detailed inferences about heterogeneous effects in an interpretable framework. This helps examine theories that predict how an effect varies and compare sources of variation.[7] Partial pooling also facilitates reasonable estimates for small groups by sharing information across groups and incorporating predictors in the heterogeneous effects equation. Finally, this approach will be faster than machine learning approaches for many datasets, easier to use in small datasets, and may scale better than models of individual treatment effects.

Like all methods, the hierarchical approach has downsides, some of which can be ameliorated with modifications, while others should lead researchers to use different tools. Because groups are based on unique combinations of heterogeneous effect variables, using multiple continuous variables in the heterogeneous effects equation creates many small groups or individual treatment effects, which increases the risk of sampling problems, especially in small datasets. If using continuous variables hinders model convergence, researchers can bin continuous variables.

Furthermore, hierarchical models can show general trends, but will not make powerful comparisons between every group. Researchers who want to compare specific groups may lack empirical leverage, especially for small groups.

With these considerations in mind, when should researchers use hierarchical models in place of interactions? If only one factor modifies an effect, interactions are best, as the extra information hierarchical models provide is less valuable.

With two or more modifiers, hierarchical models begin to add value beyond interactions. Interpreting triple interactions between a variable and two modifiers is challenging. The advantages of hierarchical modeling increase with the number of modifiers, until additional modifiers create small groups that complicate model fitting. The thresholds where the number of

[7] Rescaling variables in the heterogeneous effects equation can aid model fitting and coefficient comparisons (Gelman, 2008).

	Hierarchical Models	Interactions/Subgroup	Machine Learning
Factors	Two or more	One or two	Many
Sample Size	Conditional on number of factors	Medium to large, depending on main effect size	Large
Complexity	Medium	Low	High
Computational Cost	Medium	Low	High
Interpretability	High	High	Low
Modifiers	Specified	Specified	Discovered or Specified

Table 1. *Key characteristics of different approaches to estimating heterogeneous effects.*

modifiers becomes an issue for hierarchical modeling depends on the data, as larger datasets can support more groups.

The relative use cases of hierarchical models and machine learning are different. Unlike machine learning approaches, hierarchical models will not discover high-dimensional interactions. Researchers can add flexibility with additional interactions or non-linear specifications in either level of the model, but this requires a priori specification. Therefore, if researchers want to focus on flexible discovery, not testing an argument with multiple sources of treatment heterogeneity, they should rely more on machine-learning.

In summary, researchers should continue to use interactions for single modifiers and machine learning to discover complex interactions. Hierarchical modeling works well when there are two or more modifiers and researchers have adequate data to support an informative model. Table 1 summarizes some relevant characteristics of hierarchical, interaction and machine learning approaches to heterogeneous effects. Hierarchical modeling is thus an intermediate tool between interactions and machine-learning, where researchers need more flexibility than interactions but are not willing or able to tackle the computational and interpretation challenges of machine learning.

4 Example Application

In the following, I demonstrate how the hierarchical approach works by reanalyzing a study by Tomz and Weeks (2021) (TW hereafter). TW examine whether the public is more willing to go to war for an allied country. In a factorial experiment with vignettes, they find a 33% average increase in support for military intervention on behalf of another country if that country is an ally. This is a large and potentially important effect, so I estimate who responds to alliances.[8]

I used race, gender, hawkishness and internationalism to define demographic groups and predict treatment heterogeneity across those groups. I selected these variables because foreign policy dispositions like militant assertiveness shape general willingness to use force (Kertzer et al., 2014) as do gender (Barnhart et al., 2020) and race. I also control for other experimental manipulations.[9] Following TW's OLS analysis, I use a Gaussian likelihood, although the outcome is a binary variable.

I describe the results in two steps. First, I summarize the predictors of the alliance effect in Figure 1. I then present the resulting heterogeneous effects for every group in Figure 3.

Figure 1 plots how support for international engagement, willingness to use force, race and gender modify the impact of alliances.[10] These group-level predictors modify the treatment effect. Each coefficient therefore expresses whether a variable increases, decreases, or has no relationship with the impact of alliances on support for military intervention.

The top coefficient estimate, "Alliance" indicates that when all group variables are 0, alliances increase support for intervention by 20%. The white race coefficient suggests that alliances increase support for intervention by 12% more among white respondents than non-white respondents. Next, men are marginally more responsive to alliances than women, but

[8]See the appendix for a heterogeneous treatments analysis that corroborates TW's results.

[9]See the appendix for priors.

[10]These are the λ parameters above.

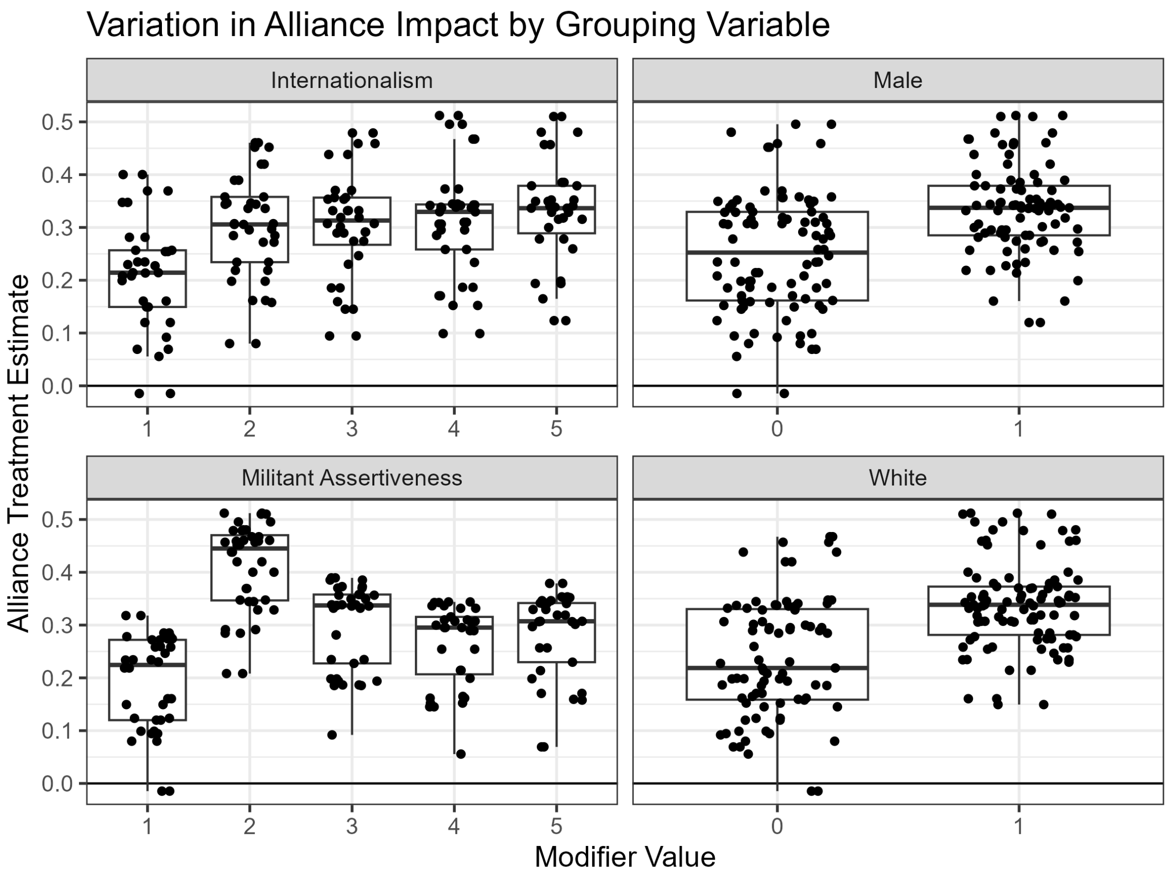

Figure 1. *Heterogeneous effects equation coefficients from a hierarchical model of how military alliances impact public support for war. Hawkishness, internationalism, white race and male gender predict the impact of alliances.*

this difference has positive, negative and zero values. As internationalism increases, the impact of alliances rises by 4% in expectation. Conversely, greater hawkishness marginally attenuates the impact of an alliance.

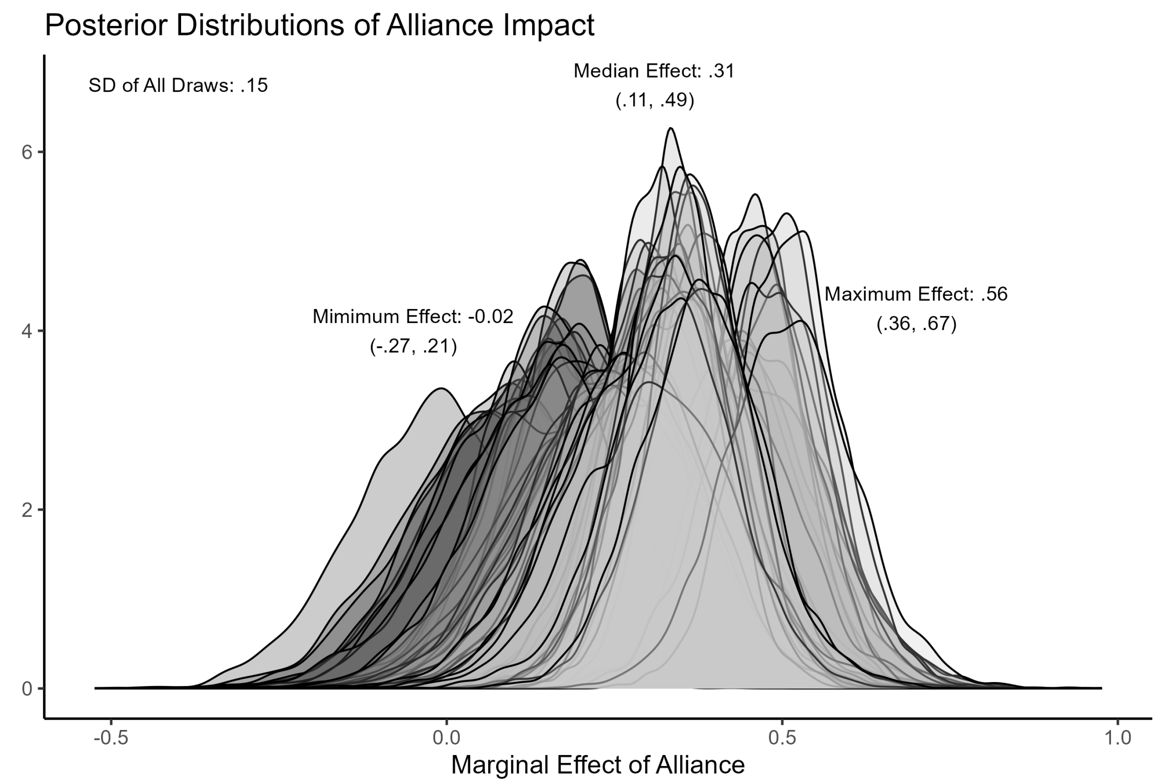

Figure 2. *Posterior distribution of all estimated impacts of alliances on support for using force. Text values give notable point estimates, and parentheses summarize the 95% credible interval.*

As a result of random variation and the systematic modification of group level variables I summarize in Figure 1, how alliances impact support for using force varies widely. Figure 2 provides an initial summary of that variation, and highlights several noteworthy estimates.

First, Figure 2 notes that the minimum estimated impact of an alliance on a demographic group is .05, while the maximum is .53. The maximum effect occurs among white men with high internationalism and low hawkishness. The minimum effect applies to non-white women with low internationalism and high hawkishness. There is no overlap in the posteriors of these estimates. The median group treatment effect estimate is .31, and this group of respondents is non-white men with middling internationalism and hawkishness. Alliances never clearly decrease support for intervention, but how much they increase support varies widely.

Figure 2 also presents the variation in how alliances impact demographic groups The standard deviation of all posterior draws is .13. Roughly 5% of variation in the alliance effect is not explained by systematic regression components in Figure 1.[11]

Figure 3 provides an alternative presentation of the impact of alliances on support for using force that bridges Figure 1 and Figure 2. As Figure 1 and Figure 2 suggest, alliances increase support for foreign intervention most among white men, especially those with low hawkishness and high internationalism. By contrast, alliances have little impact on support for war among non-white females who are also skeptical of international engagement. Individuals with more ambivalent foreign policy views respond more typically to TW's alliance treatment.

All these estimates suggest that internationalism matters more than than hawkishness for understanding who is willing to fight for U.S. allies. Alliances may impact hawks less because these individuals support intervention regardless. Military alliances matter most to backers of international engagement who are otherwise less inclined to use force.

These results show some of the strengths and weaknesses of the hierarchical approach to heterogeneous effects.[12] A simple model based on demographic groups provides precise insights about who heeds alliances in supporting using force abroad. At the same time, because some demographic groups are small, the within-group effect estimates have substantial uncertainty and powerful comparisons between most groups is challenging. Fewer groups would have more data and less uncertainty but perhaps obscure variation across key demographic characteristics.

[11] This is σ_θ above.

[12] In the appendix, I analyze Bush and Prather (2020).

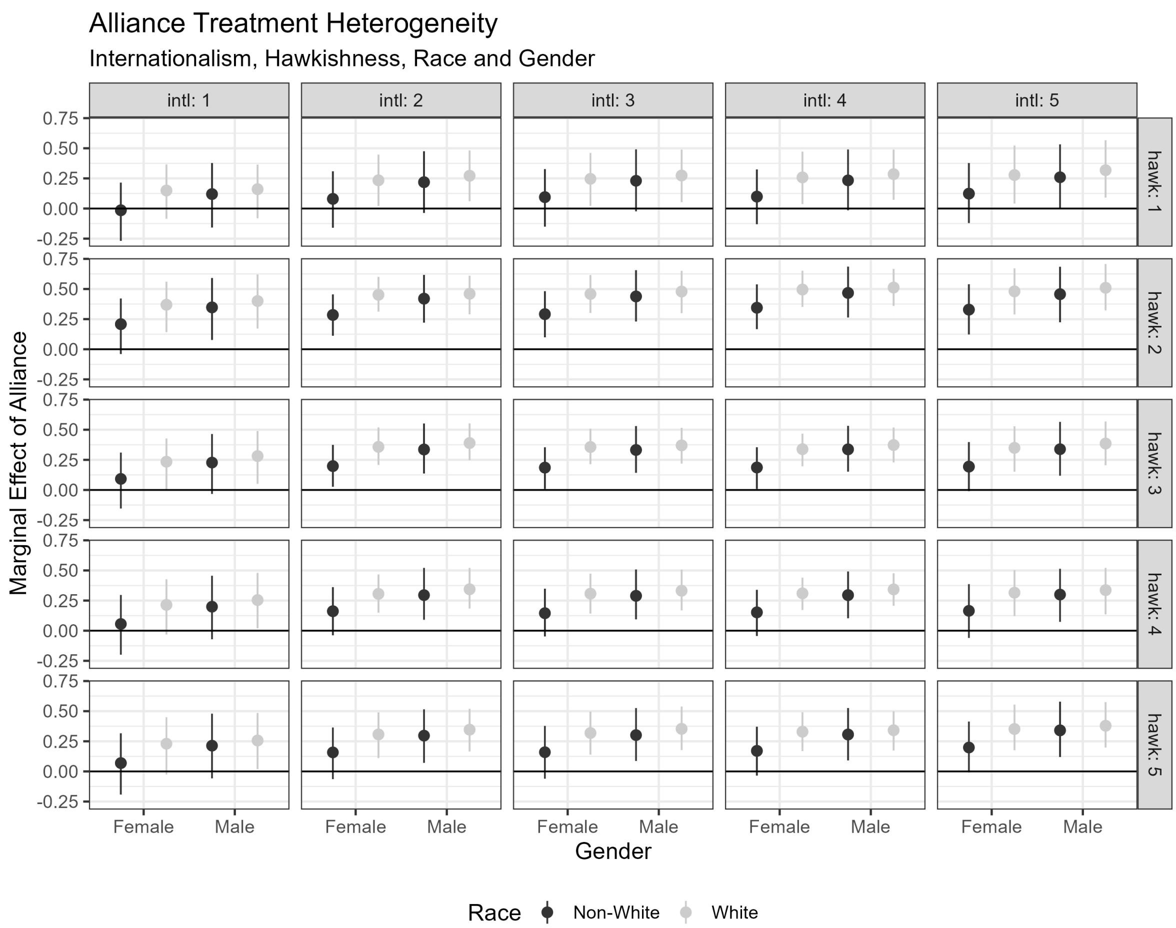

Figure 3. *Estimates of the impact of military alliances on support for using force within demographic groups. Column facets are values of internationalism, and row facets are levels of hawkishness. X-axis divided by gender and colors demarcate gender. Points mark the posterior median and bars summarize the 95% credible interval.*

5 Conclusion

This note explained how and when to use hierarchical models to estimate heterogeneous effects. Bayesian modeling can apply to a wide range of outcomes, data structures, and theories. It also details what drives variation in an effect and how much an effect varies. Explicitly modeling how different groups respond to an independent variable can help test arguments and inform policy.

Hierarchical modeling provides an intermediate approach between interactions or subgroup analyses and machine learning algorithms. For interactions with one or two variables, relying on simple interaction tools is best. Similarly, machine learning is best for discovery of complex heterogeneity. When there are two or more modifiers and many groups of theoretical interest, hierarchical modeling allows theoretically informed and interpretable estimation of effect variation.

As a result, hierarchical modeling complements existing tools and should not replace them. Researchers can use hierarchical models to check and inform other techniques, for instance by seeing if a key interaction holds when there are multiple modifiers, or comparing multiple modifiers that past theories have identified. Using hierarchical modeling can thus help scholars and policymakers better understand heterogeneous effects.

Acknowledgements

Thanks to Taylor Kinsley Chewning, Andrew Gelman and Carlisle Rainey for helpful comments.

References

Abramson, Scott F, Korhan Koçak and Asya Magazinnik. 2022. "What Do We Learn about Voter Preferences from Conjoint Experiments?" *American Journal of Political Science* 66(4):1008–1020.

Alley, Joshua. 2021. "Alliance Participation, Treaty Depth and Military Spending." *International Studies Quarterly* 65(4):929–943.

Arel-Bundock, Vincent. N.d. *marginaleffects: Predictions, Comparisons, Slopes, Marginal Means, and Hypothesis Tests*. R package version 0.14.0.9000.
URL: *https://vincentarelbundock.github.io/marginaleffects/*

Barnhart, Joslyn N, Robert F Trager, Elizabeth N Saunders and Allan Dafoe. 2020. "The Suffragist Peace." *International Organization* 74(4):633–670.

Blackwell, Matthew and Michael P Olson. 2022. "Reducing Model Misspecification and Bias in the Estimation of Interactions." *Political Analysis* 30(4):495–514.

Bürkner, Paul-Christian. 2017. "brms: An R package for Bayesian multilevel models using Stan." *Journal of Statistical Software* 80(1):1–28.

Bush, Sarah Sunn and Lauren Prather. 2020. "Foreign Meddling and Mass Attitudes Toward International Economic Engagement." *International Organization* 74(2):584–609.

Dorie, Vincent, George Perrett, Jennifer L Hill and Benjamin Goodrich. 2022. "Stan and BART for Causal Inference: Estimating Heterogeneous Treatment Effects Using the Power of Stan and the Flexibility of Machine Learning." *Entropy* 24(12):1782.

Feller, Avi and Andrew Gelman. 2015. "Hierarchical Models for Causal Effects." *Emerging Trends in the Social and Behavioral Sciences: An interdisciplinary, searchable, and linkable resource* pp. 1–16.

Gelman, Andrew. 2008. "Scaling regression inputs by dividing by two standard deviations." *Statistics in medicine* 27(15):2865–2873.

Green, Donald P and Holger L Kern. 2012. "Modeling Heterogeneous Treatment Effects in Survey Experiments with Bayesian Additive Regression Trees." *Public Opinion Quarterly* 76(3):491–511.

Grimmer, Justin, Solomon Messing and Sean J Westwood. 2017. "Estimating Heterogeneous Treatment Effects and the Effects of Heterogeneous Treatments with Ensemble Methods." *Political Analysis* 25(4):413–434.

Guisinger, Alexandra and Elizabeth N. Saunders. 2017. "Mapping the Boundaries of Elite Cues: How Elites Shape Mass Opinion across International Issues." *International Studies Quarterly* 61(2):425–441.

Imai, Kosuke and Marc Ratkovic. 2013. "Estimating treatment effect heterogeneity in randomized program evaluation." *The Annals of Applied Statistics* 7(1):443–470.

Kertzer, Joshua D., Kathleen E. Powers, Brian C. Rathbun and Ravi Iyer. 2014. "Moral Support: How Moral Values Shape Foreign Policy Attitudes." *The Journal of Politics* 76(3):825–840.

Künzel, Sören R, Jasjeet S Sekhon, Peter J Bickel and Bin Yu. 2019. "Metalearners for estimating heterogeneous treatment effects using machine learning." *Proceedings of the national academy of sciences* 116(10):4156–4165.

Marquardt, Kyle L. 2022. "Language, Ethnicity, and Separatism: Survey Results from Two Post-Soviet Regions." *British Journal of Political Science* 52(4):1831–1851.

McElreath, Richard. 2016. *Statistical Rethinking: A Bayesian course with examples in R and Stan*. CRC Press.

Simmons, Joseph P, Leif D Nelson and Uri Simonsohn. 2011. "False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant." *Psychological Science* 22(11):1359–1366.

Tomz, Michael and Jessica L.P. Weeks. 2021. "Military Alliances and Public Support for War." *International Studies Quarterly* 65(3):811–824.

Wager, Stefan and Susan Athey. 2018. "Estimation and Inference of Heterogeneous Treatment Effects using Random Forests." *Journal of the American Statistical Association* 113(523):1228–1242.