Using Hierarchical Models to Estimate Heterogeneous Effects

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

October 30, 2023

Abstract

This note describes a Bayesian hierarchical approach to estimating heterogeneous effects. To start, researchers specify groups based on quantities of interest such as heterogeneous treatments, treatment heterogeneity, and policy relevance. Then, researchers fit a hierarchical model where treatment slopes and intercepts vary across groups and group level factors modify the slopes. 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 should use hierarchical models alongside other approaches to understand heterogeneous effects for scholarship and policy.

*Thanks to Taylor Kinsley Chewning and Carlisle Rainey for helpful comments.

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, sometimes in misleading ways.[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 note describes a hierarchical Bayesian approach to estimating heterogeneous effects. There are two steps in this process. 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.[2] 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.

Such hierarchical models are easy to fit using the brms package for R (Bürkner, 2017). After fitting a model, researchers can calculate substantive effects with the marginaleffects package (Arel-Bundock, N.d.). I provide example code in this note and the appendix.

Hierarchical modeling of heterogeneous effects fills a niche between existing tools. Para-

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

[2] Using hierarchical models is an established idea in statistics (Feller and Gelman, 2015), but political science researchers rarely use them. Feller and Gelman (2015) have three applied political science citations, and of these only Marquardt (2022) models treatment effects.

metric interactions and subgroup analyses are ubiquitous because they are easy to implement and interpret. These approaches lose interpretability with more than three dimensions and are often underpowered (Simmons, Nelson and Simonsohn, 2011).[3] 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 capture complex patterns, 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 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. As a result, hierarchical modeling complements other heterogeneous effects techniques.

In the remainder of this note, I describe the approach and demonstrate how it works by analyzing a study of how military alliances shape public support for war by Tomz and Weeks (2021). The reanalysis also reveals that alliances increase support for intervention most among men who support international engagement but are otherwise skeptical of using force.

2 A Hierarchical Model of Heterogeneous Effects

There are two steps in hierarchical heterogeneous effects estimation. First, researchers must define the groups over which an independent variable's impact varies. Groups are based on unique combinations of characteristics such as other treatments, context and demographics.

Researchers should create groups based on what variation is most important and interesting.

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

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 analyses.[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 results estimate heterogeneous treatment effects. For instance, if researchers want to know how different issues shape the impact of elite foreign policy cues (Guisinger and Saunders, 2017), they could define groups by issues.

A second approach uses unit, demographic and contextual factors to create groups and estimate effect heterogeneity. In this instance, 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.

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. Using only one factor will create an unidentified model. The exact number of factors will thus depend on theory and data constraints.

After defining groups, the second step is fitting a hierarchical model of effects within

[4]It also facilitates pre-registration when applicable.

groups.[5] The first equation links the independent variable and outcome. The second equation estimates heterogeneous effects as a function of the group characteristics.[6]

This model can apply to many 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. For simplicity, assume that the outcome variable y is normally distributed with mean μ_i and standard deviation σ.[7] g indexes the researcher-defined groups.

The outcome for each unit is then a function of varying intercepts α_g, an optional matrix of control variables $\mathbf{X}$,[8] 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 predictors of heterogeneous effects $\mathbf{Z}$. Each θ parameter estimates the treatment effect in group g.

$$
\begin{aligned}
& y_i \sim N(\mu_i, \sigma) && \text{(Likelihood)} \\
& \mu_i = \alpha + \alpha_g + \theta_g T + \mathbf{X}\beta && \text{(Outcome Equation)} \\
& \theta_g \sim N(\eta_g, \sigma_\theta) && \\
& \eta_g = \lambda_0 + \mathbf{Z}\lambda && \text{(Heterogeneous Effects)}
\end{aligned}
\tag{1}
$$

The second equation then predicts the treatment effects with the matrix $\mathbf{Z}$, which contains unique combinations of whatever variables define the groups. As a result, each θ reflects a unique mix of factors that modify the treatment. The second equation also includes an intercept λ_0 that estimates the impact of treatment when all sources of heterogeneity are zero.[9]

Modeling heterogeneous effects across groups facilitates detailed inferences about when

[5]Bayesian estimation is easiest. Priors depend on the problem and researcher knowledge.

[6]Researchers can adjust for autocorrelation and clustering as needed.

[7]Researchers should use binary, categorical and other outcome likelihoods as needed.

[8]Adding additional grouping structures for more complex data is also straightforward.

[9]In brms for a model with no controls and two variables modifying the impact of a treatment, the model formula is simply y $\sim$ 1 + treat*(var1 + var2) + (1 + treat | var1:var2). treat*(var1 + var2) expresses part of the second equation, while (1 + treat | var1:var2) lets slopes vary by group.

and how much an effect varies. First, the θ parameters estimate the impact of a variable within each group.[10] All θs reflect a systematic component from the predictors in $\mathbf{Z}\lambda$ and a random component of varying slopes from σ_θ. The systematic component will usually dominate.

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 posterior distributions of different θs. The σ_θ parameter summarizes the random variation. Other techniques such as OLS with robust standard errors provide far less information.

Estimating heterogeneous effects in this way has three advantages. First, researchers can make detailed inferences about heterogeneous effects in an interpretable framework. Researchers can thus examine theories of heterogeneous effects and compare sources of variation.[11] Partial pooling also facilitates reasonable estimates for small groups by sharing information across groups and leveraging 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 that attempt to estimate individual treatment effects.

Like all methods, the hierarchical approach has downsides, some of which can be ameliorated by modifying the above framework. 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.

[10]The random intercepts α_g and varying slopes θ_g should usually have a common multivariate normal prior to capture correlations between group slopes and intercepts.

[11]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	Medium	High	Low
Modifiers	Specified	Specified	Discovered or Specified

Table 1. *Costs, benefits and key characteristics of different approaches to estimating heterogeneous effects.*

Furthermore, unlike machine learning approaches, this model will not uncover high-dimensional interactions. Even so, researchers can add flexibility with additional interactions or non-linear specifications in either level of the model. Finally, this model 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 if the groups are small. Table 1 summarizes the characteristics of hierarchical, interaction and machine learning approaches to heterogeneous effects.

3 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 how military alliances shape public support for war. 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. I estimate how demographics drive treatment heterogeneity.[12]

I used race, gender, hawkishness and internationalism to define the groups and predict

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

the impact of alliances on support for using force. 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.

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 2.

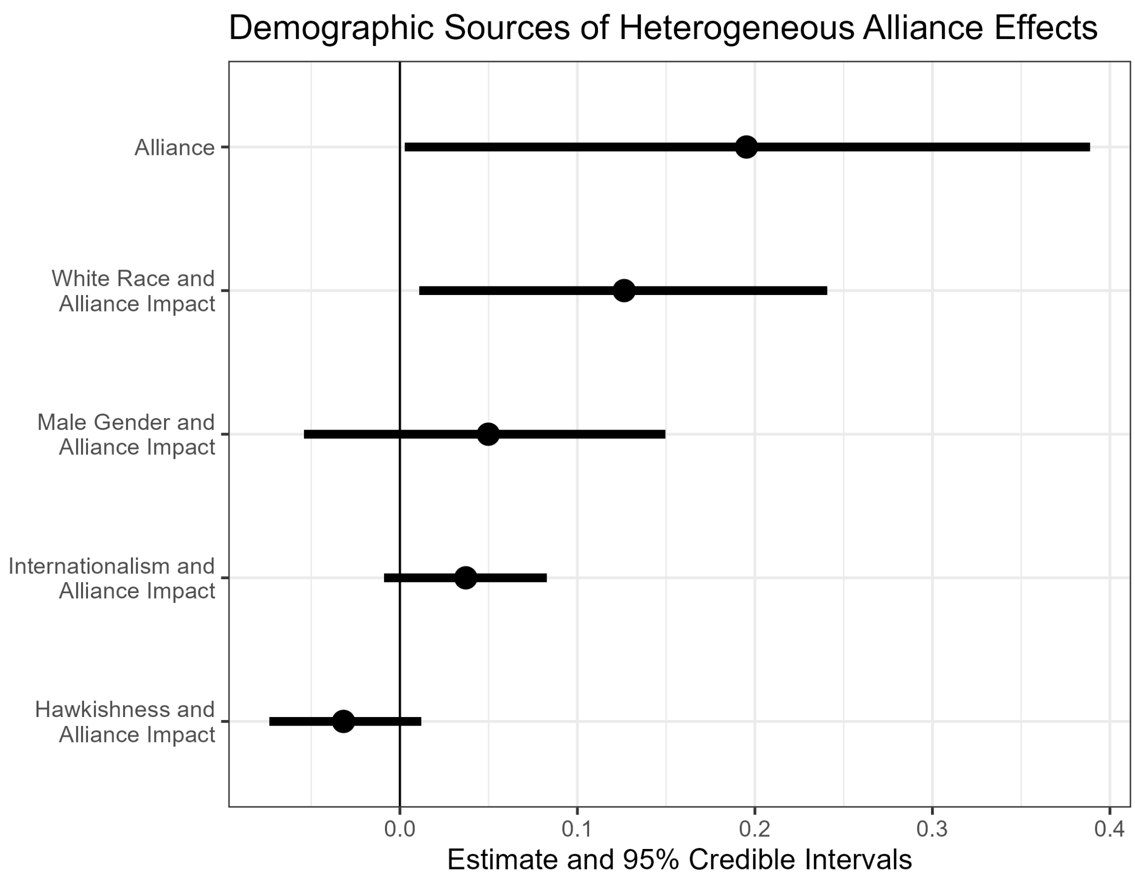

Figure 1. *Heterogeneous effects equation coefficients from a hierarchical model of how military alliances impact public support for war. Hawkishness, internationalism, white race and male gender predict the impact of alliances.*

Figure 1 plots how support for international engagement, willingness to use force, race and gender modify the impact of alliances.[13] When all these grouping variables are 0, alliances in–

[13] These are the λ parameters above.

crease support for intervention by 20%. That impact is 12% greater among white respondents. As internationalism increases, the impact of alliances rises by 4% in expectation. Greater hawkishness marginally attenuates the impact of an alliance. Furthermore, there is an additional 5% of variation in the alliance impact that these systematic components do not explain.[14]

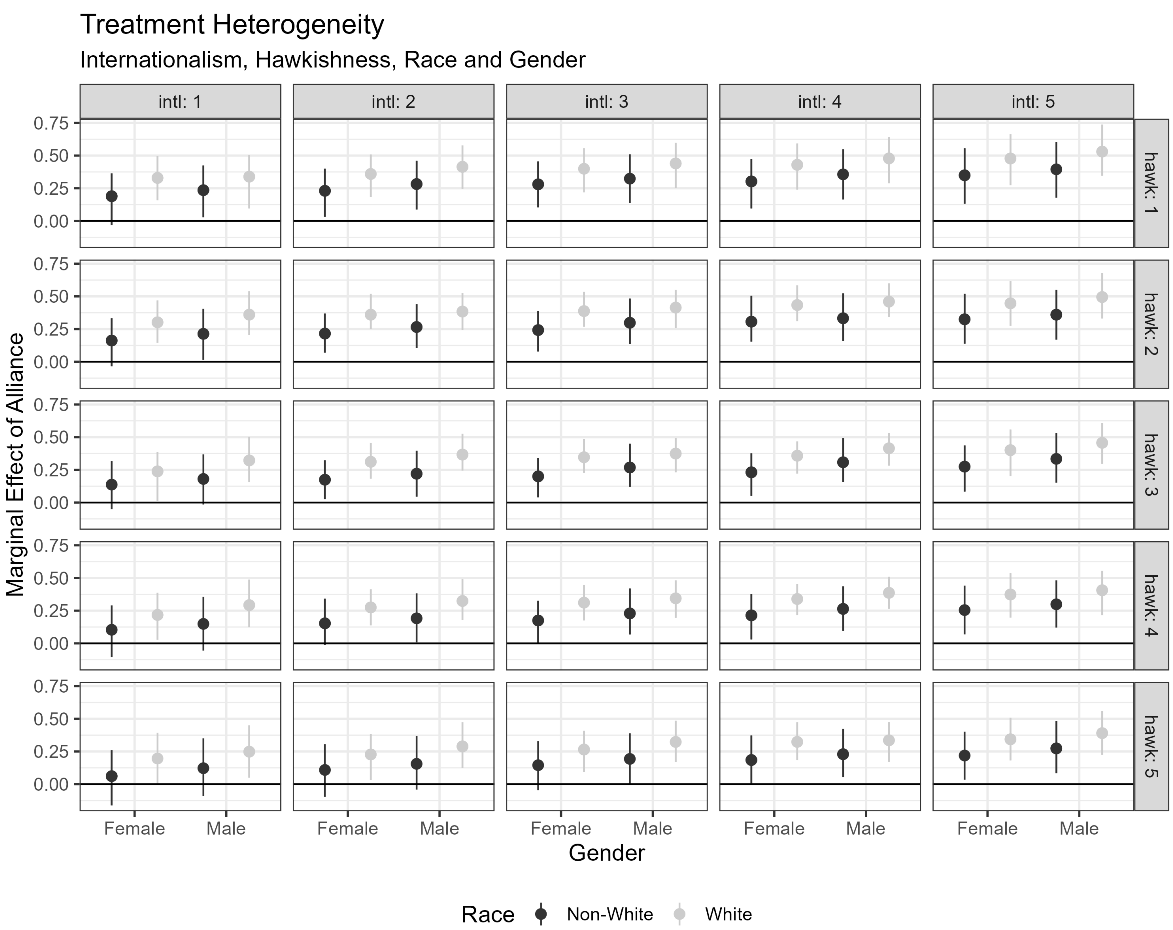

Figure 2. *Estimates of how the impact of military alliances on support for using force varies across different demographic groups. Points mark the posterior median and bars summarize the 95% credible interval.*

Figure 2 shows that alliances exert the most influence on support for foreign interventions among white men, especially those with low hawkishness and high internationalism, who can be labeled as "cooperative internationalists." Among white men with minimum hawkishness

[14]This is σ_θ above.

and maximum internationalism, alliances increase support for using force by 50%, which is roughly double the typical effect. By contrast, alliances have little impact on support for war among non-white females who are skeptical of international engagement. Militant assertiveness reduces the impact of alliances, perhaps because these individuals support intervention regardless. This implies that alliances help convince individuals who back international engagement but are less inclined to use force. As a result, internationalism is more important than hawkishness for understanding who is willing to fight for U.S. allies.

How much does the impact of alliances vary? The minimum impact of alliances is .06, and the maximum is .53, and median is .3. The standard deviation of the impact of alliances is .09. Alliances never decrease support for intervention, but how much they increase support varies widely across demographic groups.

These results show some of the strengths and weaknesses of the hierarchical approach to heterogeneous effects.[15] A simple model based on demographic groups provides new insights about who responds to alliances-. At the same time, because some demographic groups are small, the within-group effect estimates have substantial uncertainty, so comparing groups is challenging. Smaller groups would have less uncertainty but perhaps obscure variation in the impact of alliances.

4 Conclusion

This note introduced a simple and interpretable hierarchical technique for estimating heterogeneous effects. The approach above can apply to a wide range of outcomes, data structures, and theories. Explicitly modeling how different groups respond to an independent variable can help test arguments and inform policy.

Hierarchical modeling provides an intermediate approach between simple interactions or

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

subgroup analyses and complex machine-learning algorithms. It also details what drives variation in an effect and how much an effect varies. As a result, this technique 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. With this and other tools, scholars and policymakers can better understand heterogeneous effects.

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.

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.