# Estimating Treatment Effect Heterogeneity with Varying Coefficients and Bayesian Tree Ensembles

Saloni Bhogale[*] Sameer K. Deshpande[†]

March 30, 2025

**Abstract**

Understanding treatment effect heterogeneity is important, but many standard approaches to estimate heterogeneous treatment effects rely on strong assumptions. We present a novel approach to estimate heterogeneous treatment effects in a variety of experimental and observational set-ups commonly used in economics and political science such as multi-arm RCTs and difference-in-differences designs. We set up the estimation and inference problem in the form of a linear varying coefficient model and use Bayesian Additive Regression Trees (BART) to estimate the coefficients. We demonstrate how the approach can be leveraged to understand heterogeneity in multi-arm experiments and difference-in-differences designs. Our approach makes no parametric assumptions about the treatment effect function and estimates it using a flexible regression tree ensemble. We illustrate the benefits of using the approach to estimate heterogeneous treatment effects for studies across various sub-fields of political science. In doing so, our study adds to the growing literature on using flexible machine-learning methods for causal inference.

## 1 Introduction

Understanding heterogeneity of treatment effects — how the effects of interventions varies across the population — is important for social science research. Uncovering heterogeneous treatment effects not only helps scholars better understand the underlying mechanisms driving observed outcomes, but also help practitioners design personalized and effective interventions by helping them efficiently target scarce resources. To probe for potential effect heterogeneity, researchers often first fit a linear model, regressing the outcome onto treatment variables and other covariates. Then, they elaborate the initial model with additional interactions between the treatment variables and covariates, which may be discrete (e.g., indicators encoding sex or political affiliation) or continuous (e.g., age). Although fitting such models is straightforward and exceedingly common in practice, including a single multiplicative interaction implicitly assumes that the treatment's effect changes *linearly* with respect to the covariate. Such a strong functional form assumption is typically unrealistic in practice (Hainmueller et al., 2019) and the resulting model may return biased causal effects due to model misspecification (Blackwell and Olson, 2022).

One way to overcome misspecification bias is to increase the representational flexibility of the model by introducing additional interaction terms. For instance, to allow a treatment's effect to vary nonlinearly with respect to a single covariate, one may interact the treatment variable with functions of the covariate (e.g., age, age-squared, etc.) To allow a treatment's effect to vary with respect to multiple covariates, one may include higher-order interaction terms to the model. Unfortunately, when the number of covariates is large — as is increasingly the case in the empirical literature — the number of model parameter quickly explodes and the model becomes increasingly difficult to interpret. And models with many interaction terms may still be misspecified For that reason, many practitioners have turned to flexible machine learning models

[*] Department of Political Science, University of Wisconsin–Madison, bhogale@wisc.edu
[†] Department of Statistics, University of Wisconsin–Madison, sameer.deshpande@wisc.edu

to fit their data *without pre-specifying the functional form of the relationship between outcome, treatment, and covariates.* Despite their conceptual elegance and excellent predictive performance, methods relying on machine learning often require considerable more computational and statistical sophistication to implement, understand, and explain than the familiar linear models with interactions. The platonic ideal between these two extremes — misspecified but easy-to-use linear models and flexible but difficult-to-use machine learning models — would retain the familiarity and ease-of-use of the former but the expressivity of the latter.

In this paper, we demonstrate how a new Bayesian semiparametric method for fitting *varying coefficient models*, which is known as VCBART (Deshpande et al., 2024), comes close to this ideal. This article is structured as follows. In Section 2, we show how one can mitigate the risk of bias from fitting parameter models with misspecific interactions by casting the HTE estimation problem as one of fitting varying coefficient models (Hastie and Tibshirani, 1993). Then, in Section 3, we review VCBART, a new Bayesian nonparametric approach to fitting varying coefficient models, and explain how to use VCBART for both exploratory and confirmatory analyses. In Section 4, we use VCBART to reanalyze data from two recent studies that utilized different observational and experimental designs. We find substantively and statistically similar results when we estimate the ATE using the VCBART approach. Moreover, we are able to leverage sources of heterogeneity in confirmatory analysis testing known sub-groups (based on rurality Section 4.1 and ideology Section 4.2) and further uncover additional sources of heterogeneity using exploratory analysis. We conclude with a discussion in Section 5.

# 2 Motivating the use of varying coefficient models

We first review the core issues with interacted linear models in simplest setting involving a randomized control trial with a binary treatment. The issues we present, however, extend to more general settings and more involved designs. Then, we introduce varying coefficient models and explain how they generalize linear models with interactions.

## 2.1 Limitations of interaction terms

Suppose that we are interested in how the effect of a binary treatment $Z$ on an outcome $Y$ might vary with respect to a single covariate $X$. Researchers usually quantify heterogeneous effects with the *conditional average treatment effect* function ($\text{CATE}(x)$), which is the amount by which an outcome $Y$ observed under treatment ($Z = 1$) differs from the outcome observed under control ($Z = 0$), averaged across the subpopulation with covariate $x$. They usually formalize this comparison using potential outcomes (Splawa-Neyman et al., 1990; Rubin, 1974). Let $Y(1)$ and $Y(0)$ respectively denote the potential outcomes that would be observed if someone were in the exposed or control group. Formally, the target estimand is $\text{CATE}(\boldsymbol{x}) = \mathbb{E}[Y(1) - Y(0)|\boldsymbol{X} = \boldsymbol{x}]$.

Implicit in this notation is the "stable unit treatment value assumption" (SUTVA; Rubin, 1980; Imbens and Rubin, 2015). Under SUTVA and the standard (and frequently invoked) assumptions of strong ignorability and overlap, one can identify

$$\text{CATE}(x) = \mathbb{E}[Y|Z = 1, X = x] - \mathbb{E}[Y|Z = 0, X = x].$$

Thus, to estimate $\text{CATE}(x)$, it suffices to regress the observed outcomes onto the observed covariate and treatment indicator.

Perhaps the most common approach begins with the linear model

$$Y = \beta_0 + \beta_1 Z + \beta_2 X + \epsilon, \tag{1}$$

where the error $\epsilon$ is assumed to be mean-zero. Although this model is simple to fit (e.g., using ordinary least squares) and features easy-to-interpret parameters, it rigidly assumes that the treatment has a constant effect. Specifically, under the standard identification assumptions and the *model* assumption in (1), we compute CATE$(x) = \beta_1$ for all $x$.

Because simply linearly regressing $Y$ onto $X$ and $Z$ provides little information about potential effect heterogeneity, it is exceedingly common to elaborate the model in Equation (1) with a multiplicative interaction term and to assume

$$Y = \beta_0 + \beta_1 Z + \beta_2 X + \beta_3 X \times Z + \epsilon. \tag{2}$$

Implicit in (2) is the assumption that $Z$'s effect on $Y$ is *linear* in $X$. Specifically, we compute CATE$(x) = \beta_1 + \beta_3 x$. When $X$ is binary — that is, when $X$ indicates membership into one or two subgroups — this assumption is fairly innocuous, as $\beta_3$ simply measures difference in the treatment's effect in each subgroup. But when $X$ is continuous, this strong functional form assumption severely limits our ability to detect important heterogeneity and is often violated in practice (Hainmueller et al., 2019). For instance, suppose we wished to understand how the effectiveness of a voter education program changes with age. Because it assumes that effectiveness increases or decreases at a constant rate over time, the model in (2) cannot detect whether the effect attenuates or levels off as people get older. In other words, the model may be misspecified and the resulting model misspecification may result in highly biased estimates of CATE$(x)$ (see, e.g., §2.2 Blackwell and Olson, 2022).

To mitigate the risks of potential misspecification bias, it is tempting to refine model (2) by adding non-linear functions of $X$ and interactions between those non-linearities and $Z$. Specifically, for a fixed basis of pre-specified functions $\phi_1(x), \ldots, \phi_D(x)$, we could fit

$$Y = \beta_0 + \beta_1 Z + \sum_{d=1}^{D} \gamma_d \phi_d(X) + \sum_{d=1}^{D} \delta_d \phi_d(X) \times Z + \epsilon \tag{3}$$

Compared to Equations (1) and (2), the model in Equation (3) is conceptually much more compelling because it allows the treatment's effect vary non-linearly with $X$. Indeed under (3), we have CATE$(x) = \beta_1 + \sum_d \delta_d \phi_d(x)$. But fitting the more flexible model in practice is requires us to make several modeling decisions (viz., what and how many basis functions), each of which exposes us to potential misspecification bias. Returning to the voter education example, a model that interacts treatment with polynomials of age would return very biased estimates of an effect with diminishing (i.e. logarithmic) returns.

So even in the simplest setting with binary treatment and a single covariate, it is exceedingly difficult to correctly specify a sufficiently flexible linear model with interactions. These challenges are only magnified in the presence of multiple covariates and treatment variables. In Section 4.2, for instance, we re-analyze a study with $n = 2040$ observations and $p = 105$ covariates, of which 100 are binary and 5 are continuous. Including all two- and three-way interactions between binary covariates and the treatment variable introduces a huge number of parameters. Including further interactions in attempt to avoid what Blackwell and Olson (2022) term "omitted interaction bias" explodes the number of parameters. Fitting such highly parametrized models in practice involves requires careful regularization to prevent over-fitting and considerable statistical sophistication to derive standard errors under suitable asymptotic and regularity conditions.

## 2.2 Varying coefficient models

Given the nigh impossibility in correctly specifying sufficiently flexible linear models with interactions, researchers are increasingly turning to machine learning models. At a high level, most of these ML-based workflows for heterogeneous treatment effect estimation begin by first estimating the response surface $\mathbb{E}[Y|\boldsymbol{X} = \boldsymbol{x}, Z = z]$ as a function of a (potentially high-dimensional) *vector* of covariates $\boldsymbol{X}$. Typically, this estimation is done with by training a highly parametrized model like a neural network or regression tree ensemble with regularization to prevent over-fitting. Then, one computes estimates of the CATE$(\boldsymbol{x})$ by evaluating the estimated response surface at different $(\boldsymbol{x}, z)$ values and taking differences; see, for instance Hill (2011) and Athey and Wager (2019).

Although the regularization utilized in the first step of these workflows can produce extremely accurate estimates of the response surface and predictions of the outcome, they can *indirectly* yield highly biased estimates of causal estimands. Hahn et al. (2018) termed this "regularization induced confounding" while Blackwell and Olson (2022) distinguish between "direct regularization bias" and "indirect regularization bias." Further, the models used to estimate the response surface are often difficult to interpret and can demand considerable computational sophistication on the part of practitioners given the complex parameter tuning involved and lack of user-friendly software packages.

We argue that *varying coefficient* models (Hastie and Tibshirani, 1993) represent a compelling middle ground between highly parametrized (but likely misspecified) linear models with interactions and difficult-to-understand machine learning models. In its most general form, a linear varying coefficient model posits a linear relationship between an outcome $Y$ and $R$ predictors $Z_1, \ldots, Z_R$ that is allowed to change according to the value of $p$ covariates $X_1, \ldots, X_p$. Collecting the covariates into a single vector $\boldsymbol{X} = (X_1, \ldots, X_p)^\top$, the model asserts

$$Y = \beta_0(\boldsymbol{X}) + \beta_1(\boldsymbol{X})Z_1 + \cdots + \beta_R(\boldsymbol{X})Z_R + \epsilon, \tag{4}$$

where $\beta_0(\boldsymbol{X}), \ldots, \beta_R(\boldsymbol{X})$ are *functions* of a $p$-dimensional input and the error $\epsilon$ has mean-zero.

The varying coefficient model in Equation (4) generalizes the linear models with interactions in Equations (2) and (3): although the outcome is linear in $Z$, the slopes are allowed to change as arbitrary functions of the vector of covariates $\boldsymbol{X}$. In this way, varying coefficient models strike a balance between fully parametric models, which making strong functional form assumptions that are almost certainly misspecified, and fully nonparametric machine learning models, which obscure the relationship between the outcome, covariates, and treatment variables. As we detail in Section 4, we can cast the problem of heterogeneous effect estimation in a wide variety of designs popular in political science within a varying coefficients framework.

Varying coefficient models have enjoyed a long history in Statistics and Econometrics. They have also been used sparingly in the Political Science. An important early example is Jackson (1991), which not only provides examples of models of political science theories that can benefit from varying coefficients approach, but also contrasts the flexible methods with the biased OLS estimates using simulation studies. More recently, Hainmueller et al. (2019, §4.2) suggested the use of varying coefficients as an alternative to linear models with interactions. These papers, however, focused on settings with scalar covariates.

For our purposes, the ideal estimation method (i) imposes no functional form assumptions about the coefficient functions $\beta_j(\boldsymbol{x})$; (ii) returns accurate point estimates of and well-calibrated uncertainty intervals for both individual $\beta_j(\boldsymbol{x})$ values (i.e., causal quantities) and the response surface $\mathbb{E}[Y|\boldsymbol{X} = \boldsymbol{x}, \boldsymbol{Z} = \boldsymbol{z}]$; (iii) scales to datasets with many observations (i.e., large $n$) and covariates (i.e., large $p$); and (iv) can be run easily "off-the-shelf" by researchers who do not have considerable statistical or computational expertise.

Unfortunately, until recently, most procedures for fitting varying coefficients with $p > 1$ covariates fail to meet all these desiderata. For instance, Lee et al. (2012) and Tibshirani and Friedman (2020) both assumed that the $\beta_j(\boldsymbol{x})$'s additive functions of the covariates. Although Li and Racine (2010)'s kernel smoothing procedure does not impose strong structural assumption, it involves tuning a several bandwidth parameters with leave-one-out cross-validation, which becomes computationally prohibitive when $n$ and $p$ are large. Further, their implementation, which is available in the **np** R package (Hayfield and Racine, 2008), does not return uncertainty interval by default, forcing practitioners to roll their own re-sampling or bootstrap procedures. In Section 3 we introduce a relatively new Bayesian nonparametric method, VCBART (Deshpande et al., 2024), which meets our desiderata.

# 3 Estimating varying coefficient models using VCBART

Deshpande et al. (2024) introduced a new Bayesian semiparametric model called VCBART to estimate varying coefficient models. At a high level, VCBART works by approximating each function $\beta_j(\boldsymbol{x})$ in Equation (4) with an ensemble of regression trees (i.e., piecewise-constant step functions). As a Bayesian procedure, VCBART formally computes a posterior distribution over the collection of regression tree ensembles. It is this posterior — and not hypothesized repeated sampling — that serves as the basis of all of our inference about heterogeneous individual and group-level causal effects. Because the posterior is analytically intractable, posterior summaries are computing using Markov chain Monte Carlo (MCMC).

In Section 3.1, we provide a high-level overview of the VCBART prior and describe the MCMC computationally strategy; we refer interested readers to §3 and Appendix S4 of Deshpande et al. (2024) for the full technical details. Then, in Section 3.2, we describe how to use the output of VCBART to draw

## 3.1 The VCBART prior & Gibbs sampler

Before describing how VCBART works, we introduce some notation. Without loss of generality, suppose that all covariates have been scaled to the interval $[0, 1]$ (i.e., $\boldsymbol{X} \in [0, 1]^p$) [1]. Formally, a *regression tree* is the pair $(\mathcal{T}, \Lambda)$ consisting of (i) a finite binary decision tree $\mathcal{T}$ containing several terminal (or leaf) nodes and several non-terminal or decision nodes; and (ii) a collection $\Lambda$ of scalars, one for each terminal or *leaf* node of the tree. Each decision node of $\mathcal{T}$ is associated with a decision rule of the form $\{X_j < c\}$. Given a decision tree $\mathcal{T}$, we can associate every $\boldsymbol{x} \in [0, 1]^p$ with a unique leaf node, which we denote $\ell(\boldsymbol{x}; \mathcal{T})$ by tracing a path from the root of $\mathcal{T}$ down the tree. Starting from the root node, whenever the path encounters the rule $\{X_j < c\}$, the path proceeds to left child if $x_j < c$ and to the right otherwise until it reaches a leaf node, which we denote $\ell(\boldsymbol{x}; \mathcal{T})$. In this way, the decision tree $\mathcal{T}$ partitions $[0, 1]^p$ into several disjoint axis-parallel rectangular boxes, one for each leaf node. By associating each leaf $\ell$ with its own scalar $\lambda_\ell$ and denoting the collection of all $\lambda_\ell$'s by $\Lambda$, the pair $(\mathcal{T}, \Lambda)$ represents a piecewise constant function over $[0, 1]^p$ (Figure 1). Formally, we let $g(\boldsymbol{x}; \mathcal{T}, \Lambda) = \lambda_{\ell(\boldsymbol{x};\mathcal{T})}$ denote the evaluation function that returns the scalar in $\Lambda$ associated to the leaf $\ell(\boldsymbol{x}; \mathcal{T})$.

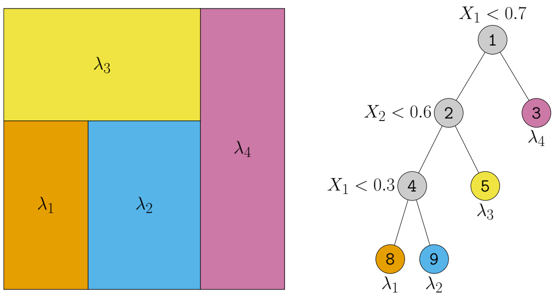

Figure 1: Example of a piecewise constant step function defined over $[0, 1]^2$ (left) and its regression tree representation (right)

For each $r = 0, \ldots, R$, VCBART introduces a collection (hereafter ensemble) of $T$ regression trees $\mathcal{E}_r = \{(\mathcal{T}_{r,t}, \Lambda_{r,t})\}_{t=1}^{T}$ and approximates each $\beta_r(\boldsymbol{X})$ as a sum of the trees in $\mathcal{E}_r$

$$\beta_r(\boldsymbol{x}) = \sum_{t=1}^{T} g(\boldsymbol{x}; \mathcal{T}_{r,t}, \Lambda_{r,t}). \tag{5}$$

**Regression tree prior.** VCBART models each ensemble $\mathcal{E}_0, \ldots, \mathcal{E}_R$ independently *a priori* and specifies independent and identical priors on the individual regression trees within each ensemble. The regression tree is best described implicitly with a three-step procedure for simulating prior samples. First, the graphical structure — that is the overall arrangement of the internal and leaf nodes and the edges — is simulated

[1] VCBART can handle both continuous and categorical predictors. See Deshpande (2024) and Deshpande et al. (2024) for details.

recursively with a branching process: starting from the root node, every time a new node is added to the tree, two more child nodes are randomly attached to it with decreasing probability. Then, conditionally on the graphical structure, a decision rule is drawn for every internal node by first randomly selecting the splitting variable $X_j$ uniformly and then sampling the cutpoint $c$ uniformly from the set of valid values of $X_j$ at that node. In Figure 1, the set of valid values for $X_1$ and $X_2$ at node 5 are, respectively, $[0, 0.7]$ and $[0.6, 1]$. Finally, the scalars $\lambda_\ell$ are drawn independently from mean-zero normal distributions with small variance.

The regression tree prior introduces strong regularization towards relatively shallow trees that do not explain too much variation in $Y$ (Chipman et al., 2010). That is, the prior strongly encourages each tree to be a "weak learner," in the sense that while no individual tree in the ensemble $\mathcal{E}_r$ closely approximates the function $\beta_r(\boldsymbol{x})$, their aggregation does. Although the regression tree prior depends on several prior hyperparameters, building on recommendations from Chipman et al. (2010), Deshpande et al. (2024) provide several default choices that work well across a wide range of problems; see §3.1 and Appendix S2 of that paper for full details.

**Sampling from the VCBART posterior.** Like most extensions of BART, VCBART uses a Gibbs sampler to simulate approximate samples from the joint posterior distribution of the regression tree ensembles and $\sigma^2$. In each iteration, (i) the regression trees are updated sequentially one-at-a-time while keeping the other trees and $\sigma^2$ and then (ii) $\sigma^2$ is updated conditionally on keeping all the trees fixed. Each regression tree update involves sampling a new pair $(\mathcal{T}, \Lambda)$ from its conditional posterior given all the other trees, $\sigma^2$, and the observed data. We sample a new regression tree in two steps. First, we draw a new decision tree $\mathcal{T}$ from its marginal distribution using a Metropolis-Hasting steps in which a random proposal is made by either growing or pruning the existing tree structure. Then, new leaf node parameters $\Lambda$ are sampled *conditionally* given the new decision tree. For a full derivation of the sampler, please see §3.2 and Appendix S3 of Deshpande et al. (2024).

## 3.2 Summarizing the posterior of causal estimands

The VCBART posterior quantifies the joint uncertainty about the functions $\beta_0(\boldsymbol{X}), \ldots, \beta_R(\boldsymbol{X})$ and the residual standard deviation $\sigma$ based on the observed data and the prior. Uncertainty about these *functions* induces uncertainty about the *values* of several quantities including, but not limited to, a single function (e.g., $\beta_r(\boldsymbol{x})$) or contrasts of multiple functions (e.g., $\beta_r(\boldsymbol{x}) - \beta_s(\boldsymbol{x})$) evaluated at a specific input or averaged across multiple inputs. Under suitable identifying assumptions, such quantities capture the heterogeneous effect of a particular treatment relative to control or other treatment for a particular individual or subgroup of individuals. We can use the draws returned by the VCBART Gibbs sampler to simulate draws from the posterior distributions of these causal quantity of interest.

For the ease of exposition, we describe this process in the simplest setting with a binary treatment (i.e., $R = 1$) and under suitable assumptions to identity $\beta_1(\boldsymbol{X}) = \text{CATE}(\boldsymbol{X}) = \mathbb{E}[Y(1) - Y(0)|\boldsymbol{X}]$. Formally, consider the $m$-th posterior draw $(\mathcal{E}_0^{(m)}, \mathcal{E}_1^{(m)}, \sigma^{(m)})$. Using the sampled regression tree ensemble $\mathcal{E}_1^{(m)}$, we can compute, for all subjects $i$ in our sample, a posterior draw of the CATEevaluated at their covariates using Equation (5):

$$\beta^{(m)}(\boldsymbol{x}_i) = \sum_{t=1}^{T} g(\boldsymbol{x}_i; \mathcal{T}_{r,t}^{(m)}, \Lambda_{r,t}^{(m)}).$$

For each individual $i$, we can then approximate summaries like the posterior mean or posterior credible intervals using the samples $\beta_1^{(1)}(\boldsymbol{x}_i), \ldots, \beta_1^{(M)}(\boldsymbol{x}_i)$.

Beyond summarizing the posterior of individual CATEevaluations, we can use the samples of $\mathcal{E}_1$ to compute samples of sample average effects. For instance, we can compute the $m$-th posterior draw of the sample average treatment effect (ATE) and samples average treatment effect on the treated (ATT) as

$$\mathrm{ATE}^{(m)} = n^{-1}\sum_{i=1}^{n}\beta_1^{(m)}(\boldsymbol{x}_i)$$

$$\mathrm{ATT}^{(m)} = n_1^{-1}\sum_{i:z_i=1}\beta_1^{(m)}(\boldsymbol{x}_i),$$

where $n_1 = \#\{i : z_i = 1\}$ counts the number of treated subjects.

**Confirmatory subgroup analyses**. In some scenarios, theory may suggest that a treatment's effect differs across known subgroups. We can use the VCBART samples to formally test whether there is a difference in the average effect within two known subgroups, $\mathcal{S}_0$ and $\mathcal{S}_1$. Letting $n_0 = |\mathcal{S}_0|$ and $n_1 = |\mathcal{S}_1|$, be the number of subjects in each subgroup, we can compute the $m$-th posterior sample of the difference in average treatment effect within each subgroup is simply

$$\Delta^{(m)}(\mathcal{S}_0, \mathcal{S}_1) = n_0^{-1} \times \sum_{i\in\mathcal{S}_0}\beta_1^{(m)}(\boldsymbol{x}_i) - n_1^{-1} \times \sum_{i\in\mathcal{S}_1}\beta_1^{(m)}(\boldsymbol{x}_i).$$

We can then report the mean (resp. quantiles) of the $\Delta^{(m)}(\mathcal{S}_0, \mathcal{S}_1)$'s as a posterior point estimate and uncertainty intervals of the difference in average subgroup effect. We can further approximate the posterior probability that the treatment's effect in subgroup $\mathcal{S}_0$ exceeds its effect in $\mathcal{S}_1$ with the proportion of positive $\Delta^{(m)}(\mathcal{S}_0, \mathcal{S}_1)$'s.

Importantly, subgroup analysis in VCBART does not require one to pre-specify the subgroups in advance. Instead, one can re-use the same posterior samples of tree ensembles to generate posterior samples of average effects within as many subgroups as one likes. One could even re-use the samples to compare two subgroups based on a comparison of two other subgroups. This is in sharp contrast with the interacted linear model, in which subgroups must be pre-specified and explicit interactions between subgroup indicators and treatment must be included in the model in advance. Conducting post-hoc subgroup analyses with the interacted linear model often involves running several model specifications. It is generally difficult to perform simultaneously valid frequentist inference based on adaptively re-running multiple model specifications with the same data. Re-using the VCBART samples to conduct additional post-hoc subgroup analyses does not face the same sorts of multiple testing issues. This is because the basis of our inference remains the same posterior; we are simply computing different aspects of this single distribution.

**Exploratory subgroup analysis**. We can even use the VCBART posterior samples to *discover* and compre new subgroups. This is especially helpful in settings where $p$ is large and we lack strong theory about what might drive treatment effect heterogeneity. We do this using a three-step procedure sometimes called "fitting-the-fit", which as emerged as an increasingly popular way to summarize heterogeneity in a parsimonious fashion (see, e.g., Hahn and Carvalho, 2015; Puelz et al., 2017; Fisher et al., 2020; Bolfarine et al., 2024; Fisher et al., 2024).

In the first step, we compute the posterior mean of $\beta_1(\boldsymbol{x}_i)$ for each subject $i$. Then, we build a single classification and regression tree to predict these posterior means using the fully vector of covariates. Each node in this fitted tree represent a different subgroup defined by different combinations of covariate splits. We can then use the procedure described above to compute, for instance, the posterior probability that the effect is greater in one discovered subgroup than in another. Importantly, so long as we just re-use the same posterior samples, such adaptive subgroup discovery does not introduce multiple testing challenges.

# 4 Empirical Examples

## 4.1 Crime reporting

Sukhtankar et al. (2022) conduct experiments to evaluate reforms that improve police responsiveness to women in India. More specifically, they introduce dedicated spaces for women in local police stations staffed by trained officers and test whether it can address under-reported gender-based violence. The study finds that officers in stations with the dedicated spaces (called *Women's Help Desks* or WHDs) are more likely to register cases of gender-based violence.

The experiment consists of three arms: the control arm $Z_0$, and the two treatment arms $Z_1$ and $Z_2$ . $Z_1$ is a bundled treatment of three interventions: private spaces, officer training and outreach to local safety networks. $Z_2$ adds a fourth component to the bundle in $Z_1$ – that the WHDs would be exclusively run by female officers. The unit of randomization is at the police station, and researchers are interested in evaluating these two arms on registered cases of domestic violence (which is a substantively important subset of gender-based violence cases).

**Identification and Estimation** In this experiment, we are interested to understand how the effect of treatments $Z_1$ and $Z_2$ vary with possible covariates. We learn from the replication archives for this paper that the researchers could only make limited attributes publicly available, since the names of police stations (the unit of analysis) and accompanying characteristics were retracted to preserve anonymity as required by the IRB. Nevertheless, they share one binary covariate (rurality) and seven continuous covariates (baseline measures of violence) that form our possible p = 8 covariates $X = (X_1, ..., X_8)^T$ of interest.

Let $Y(0)$, $Y(1)$ and $Y(2)$ denote the potential outcomes under $Z_0$, $Z_1$ and $Z_2$ respectively. Under SUTVA, strong ignorability and overlap, we can identify the target estimand to be: $\text{CATE}(x) = \mathbb{E}[Y|Z_2 = 1, X = x] - \mathbb{E}[Y|Z_2 = 0, X = x]$ and $\text{CATE}(x) = \mathbb{E}[Y|Z_1 = 1, X = x] - \mathbb{E}[Y|Z_1 = 0, X = x]$. We estimate the $CATE(x)$ by estimating $\beta_j(\boldsymbol{X})$ in the varying coefficient model:

$$Y = \beta_0(\boldsymbol{X}) + \beta_1(\boldsymbol{X})Z_1 + \beta_2(\boldsymbol{X})Z_2 + \epsilon \tag{6}$$

**Main Results**. Using VCBART, we are able to estimate individual level treatment effects as seen in Figure 2 for $\beta_1(\boldsymbol{X})$ and Figure 3 for $\beta_2(\boldsymbol{X})$. From the ITE plots, we can see that there are some police stations with effects close to 0, while there are other police stations with larger effects, which hints at possible unexplored sources of heterogeneity. We compute the ATE by taking the average of the samples of the average treatment effect and find that $Z_1 = 1.4$, $Z_2 = 1.3$ as compared to the OLS results where $Z_1 = 1.5$ and $Z_2 = 1.4$.

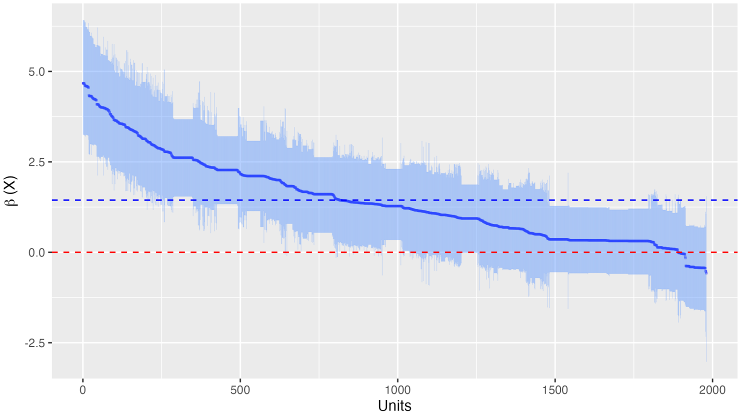

Figure 2: Individual Treatment Effects: Regular help-desks

**Confirmatory subgroup analysis based on rurality**. We can test one possible source of heterogeneity by comparing police stations in rural and urban areas. Using the process explained in Section 3.2, we create two known subgroups: $\mathcal{S}_0$ consisting of urban police stations and $\mathcal{S}_1$ consisting of rural police stations. For each posterior sample, we compute the difference of the average effect in each subgroup. We finally plot a histogram of these differences in Figure 4. Through this exercise, we learn that treatment effect is highest

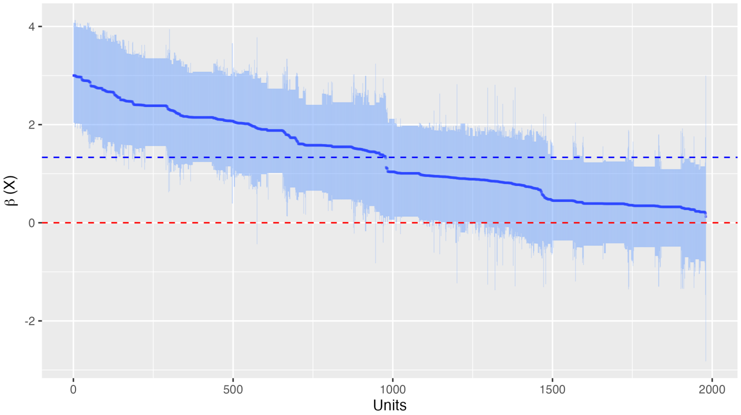

Figure 3: Individual Treatment Effects: Women run help-desks

in urban areas: for the first treatment condition, 0.89 more cases are registered per month in urban areas as compared to rural areas. This has important policy implications - given that the intervention is much more effective in urban areas, it either be efficient to roll-out the intervention to these stations first, or the intervention might need more components to be equally effective in rural areas. Importantly, we are able to say this with high confidence – the posterior probability that the treatment in $\mathcal{S}_0$ is greater than $\mathcal{S}_1$ is 1. In other words, we find that in 100% of the posterior samples, the effect for the subgroup of urban police stations is greater than the effect for the subgroup of rural police stations.

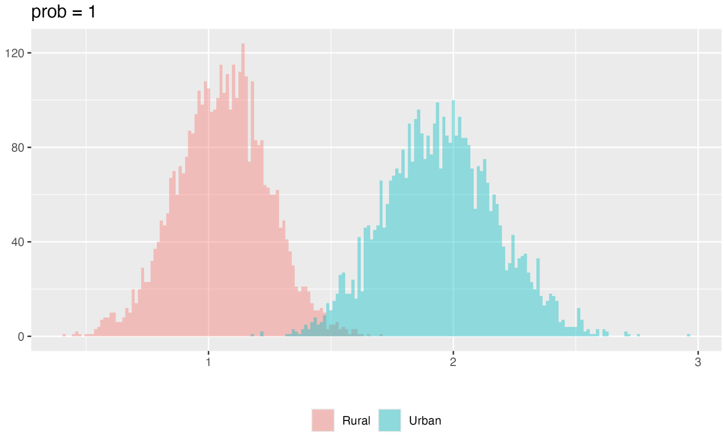

Figure 4: Comparing CATEs for Rural and Urban Police Stations

## 4.2 Narrative strategies

Exclusionary attitudes are harmful - and previous experiments have found that interpersonal conversations having multiple narrative strategies can reduce exclusionary attitudes. Kalla and Broockman (2023) attempt to distinguish between three narrative strategies that could work - essentially trying to 'unbundle' the treatment, in order to understand which of the conversational strategies actually reduce exclusionary attitudes and reduce prejudice. Understanding sources of heterogeneity can help us answer *who* is most affected by the interpersonal conversations employing various narrative strategies.

We reanalyze these three experiments and use VCBART to flexibly model the outcomes. Our contribution is to better understand heterogeneity of treatment effects of the many experiments. This can help us answer *who* is most affected by the interpersonal conversations employing various narrative strategies. In doing so, we demonstrate that our approach is suitable for a number of different design choices. We re-analyze two experiments from this paper.

The first experiment consists of two narrative strategies – respondents are randomly assigned to either a control condition $Z_0$ where canvassers have short conversations about an unrelated topic with the respondents. The first treatment arm $Z_1$ consists of three components: a 'Perspective-Getting' strategy which involves the canvasser sharing a story about an immigrant, a 'Vicarious Perspective-Giving' strategy that involves the respondent sharing a personal narrative about an undocumented immigrant and an 'Analogic Perspective-Taking' strategy that asks respondents to share a time when they needed care or support. The second treatment arm $Z_2$ is the same as intervention $Z_1$, except that the 'analogic perspective-taking' narrative strategy is excluded. The authors main outcome of interest is an index based on respondents' policy support for unauthorized immigrants in government programs and a prejudice index.

**Identification and Estimation #1** Similar to the previous experiment, we are interested to understand how the effect of treatments $Z_1$ and $Z_2$ vary with possible covariates. A rich set of covariates is captured for each respondent – a hundred binary and five continuous together form our potential p = 105 covariates $X = (X_1, X_2, ...X_{105})$ that can drive treatment effect heterogeneity.

Let $Y(0)$, $Y(1)$ and $Y(2)$ denote the potential outcomes under $Z_0$, $Z_1$ and $Z_2$ respectively. Under SUTVA, strong ignorability and overlap, we can identify the target estimand to be: $\text{CATE}(x) = \mathbb{E}[Y|Z_2 = 1, X = x] - \mathbb{E}[Y|Z_2 = 0, X = x]$ and $\text{CATE}(x) = \mathbb{E}[Y|Z_1 = 1, X = x] - \mathbb{E}[Y|Z_1 = 0, X = x]$. We estimate the $CATE(x)$ by estimating $\beta_j(\boldsymbol{X})$ in the varying coefficient model:

$$Y = \beta_0(\boldsymbol{X}) + \beta_1(\boldsymbol{X})Z_1 + \beta_2(\boldsymbol{X})Z_2 + \epsilon \tag{7}$$

**Main Results #1**. In order to estimate the ATE by using VCBART, we first set up the data in its 'long' format since each respondent responds at many time points. Next we set up the same covariates as categorical or continuous covariates and add an additional categorical variable denoting the time period in which the response was collected. Finally, we use VCBART to flexibly estimate treatment effects of the interventions on the combined index. We compute the ATE by taking the average of the sample average treatment effect associated with each posterior draw. We find that the full intervention has an ATE (with credible intervals) of 0.089 (0.06, 0.11) and the partial intervention has an ATE of 0.12 (0.08, 0.14). This is similar both in size and magnitude with the author's chosen linear model.

**Confirmatory subgroup analysis based on ideology** Attitudes towards immigration is a polarizing issue across partisan lines, thus we may expect that these effects are driven by the respondents' ideologies. We thus calculate the CATEs for three groups that are formed based on whether the respondent is conservative, liberal or moderate. In Figure 5, we show the results of this exercise. As expected, the estimate of the CATE differs by respondent's ideologies – while both types of interventions have the lowest effect on republicans the partial intervention has a slightly greater effect on the outcome for conservative respondents. Across the two interventions, the effect of the treatment appears to be greater for liberal and moderate respondents.

Further, we test these differences by computing the CATE for the subgroup in each posterior sample. Indeed, we find that for both the interventions, the differences in how conservatives and liberals or moderates respond to the the treatment is meaningful. In the first treatment arm, for over 99% of the posterior samples, the effect of the treatment for liberals or moderates is greater than the treatment effect for the conservative respondents. In the second treatment arm, conservatives continue to have lower treatment effects relative to liberals (in 99% of the posterior samples, the treatment effect for liberals is greater than the treatment effect for moderates), but the differences in conservatives and moderates is not as stark as the first intervention.

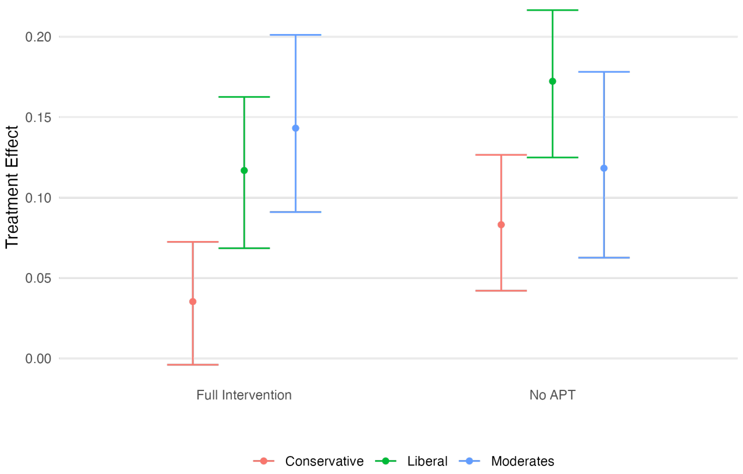

Figure 5: Narrative Strategies Effect by Ideology

**Exploratory subgroup analysis #1** We now use the VCBART posterior samples to discover new subgroups – a data-driven approach to learn about heterogeneity across different units. To do so, we provide the posterior means of $\beta_j(\boldsymbol{X})$ as inputs to fit a Classification and Regression Tree to obtain nodes and 'splits' that form sub-groups that best predict the treatment effects, and then test whether the subgroups formed are meaningfully different from each other.

We observe in Figure 7 that indeed pre-treatment measures of ideology determine the magnitude of treatment effect, as indicated by the first split based on *t0_ideology* which measures whether the respondent is liberal, moderate or conservative. Further, we learn that within the conservative group of respondents, the treatment effects for the most conservatives differ based on how self-reassured they report themselves to be. Particularly, treatment does not appear to have any effect on respondent who identify as being 'extremely conservative' *and* 'extremely confident in my abilities' (which constitute about 10% of the sample, corresponding to a treatment effect of -0.0077 as denoted in Figure 7). We test whether the sub-groups are meaningfully different from each other by computing the difference of the average effect in each subgroup. From Figure 8, we can say that in 93% of the posterior samples, the most self assured conservatives had lower treatment effects than other conservatives.

Next, we look at results from a second experiment in the same paper. In this experiment, the two treatment arms include a 'Full intervention' ($Z_1$) and the 'Perspective Getting' strategy only (where the canvasser sharing a story about an immigrant) $Z_2$.

**Identification and Estimation #2** Similar to the previous two experiments, we are interested to understand how the effect of treatments $Z_1$ and $Z_2$ vary with possible covariates. Forty covariates is captured for each respondent – of which six are continuous (age, baseline outcomes etc). Together form our potential p = 40 covariates $X = (X_1, X_2, ...X_{40})$ that can drive treatment effect heterogeneity.

Let $Y(0)$, $Y(1)$ and $Y(2)$ denote the potential outcomes under $Z_0$, $Z_1$ and $Z_2$ respectively. Under SUTVA, strong ignorability and overlap, we can identify the target estimand to be: $\text{CATE}(x) = \mathbb{E}[Y|Z_2 = 1, X = x] - \mathbb{E}[Y|Z_2 = 0, X = x]$ and $\text{CATE}(x) = \mathbb{E}[Y|Z_1 = 1, X = x] - \mathbb{E}[Y|Z_1 = 0, X = x]$. We estimate the $CATE(x)$ by estimating $\beta_j(\boldsymbol{X})$ in the varying coefficient model:

$$Y = \beta_0(\boldsymbol{X}) + \beta_1(\boldsymbol{X})Z_1 + \beta_2(\boldsymbol{X})Z_2 + \epsilon \tag{8}$$

**Main Results #2** The authors estimate treatment effects using OLS with pre-treatment covariates and

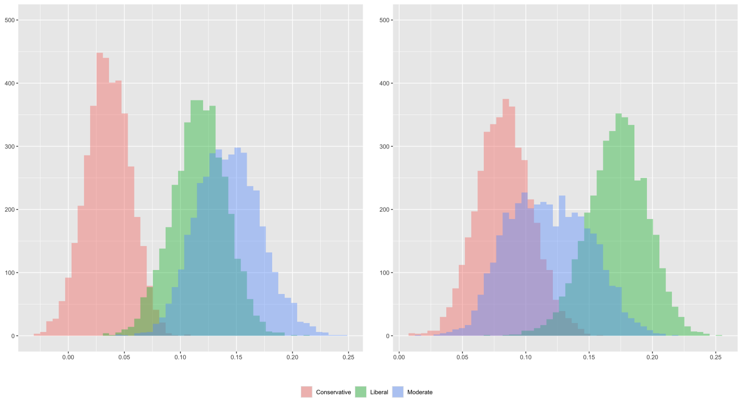

Figure 6: Narrative Strategies Effect: Comparisons among sub-groups
The figures in the first column compare sub-groups among the first arm of the treatment. The figures in the second column compare sub-groups among the second arm of the treatment.

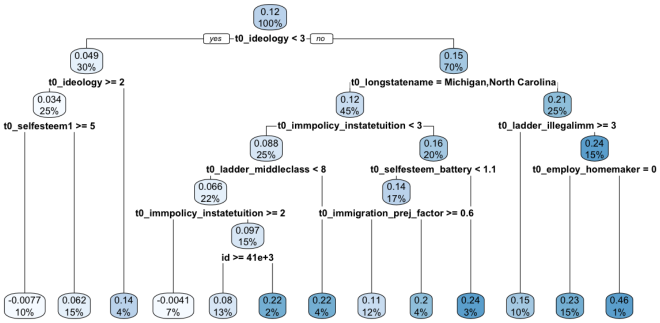

Figure 7: CART for Partial Intervention

adjust for standard errors using clustering. We compute the ATE using VCBART by taking the average of the sample average treatment effect associated with each posterior draw and find a treatment effect of 0.08 (0.04, 0.12) for the full intervention, and 0.07 (0.01, 0.13) for the perspective-getting only intervention. This is substantively similar to the results reported by the authors.

**Exploratory subgroup analysis #2**. We now use the VCBART posterior samples to discover new subgroups as before for the perspective-getting only intervention. In Figure 9, we observe that if the respondent is 65 or above in age, the treatment has a much lower effect. We use this insight to create two subgroups associated with older and younger respondents and plot the difference in the treatment effects within the subgroups in Figure 10. In all of the posterior samples, treatment effects for the older group were smaller than the treatment effects for the younger group. While the average treatment effect is 0.07, we find that it is largely driven by large effects for respondents below the age of 65.

# 5 Conclusion

Estimating heterogeneous treatment effects from varied research designs are valuable to highlight important differences in how treatment effect varies by sub-groups. However, current approaches involve invoking strong assumptions about linearity of effects and do not scale well as researchers attempt to explore multiple

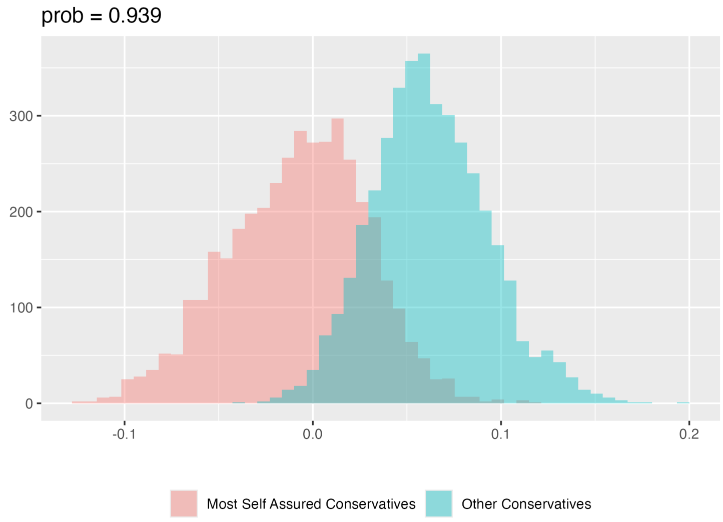

Figure 8: Conservatives: subgroups based on self-assurance

sources of heterogeneity.

In this paper, we introduce how estimating the treatment effects as flexible functions of covariates in a varying coefficient model can overcome these shortcomings. We use VCBART to estimate the covariate functions which has a number of advantages. Its ease of use and ability to discover sources of heterogeniety make it a strong candidate for use in a large number of research designs – spanning multi-arm experiments and difference-in-differences designs.

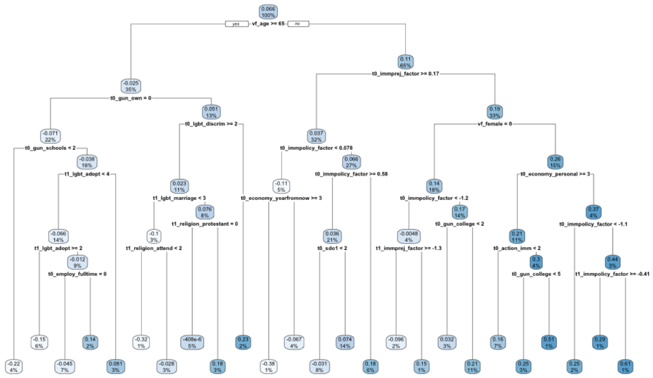

Figure 9: CART fit based on Perspective-Getting Intervention

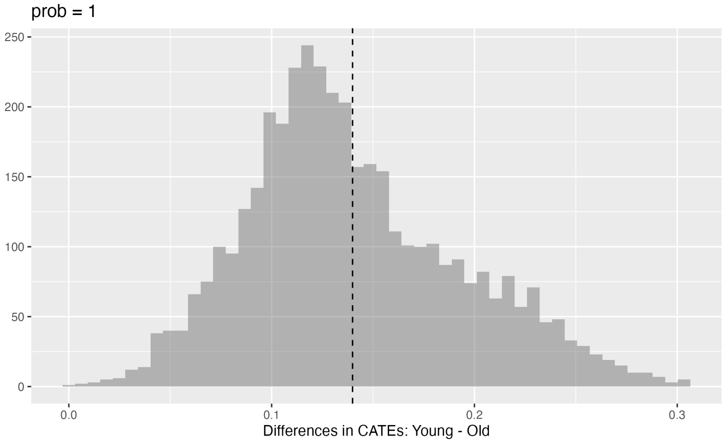

Figure 10: Testing Heterogeneity based on age

# References

Athey, S. and S. Wager (2019). Estimating treatment effects with causal forests: An application. *Observational Studies 5*(2), 37–51.

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

Bolfarine, H., C. M. Carvalho, H. F. Lopes, and J. S. Murray (2024). Decoupling shrinkage and selection in gaussian linear factor analysis. *Bayesian Analysis 19*(1), 181–203.

Chipman, H. A., E. I. George, and R. E. McCulloch (2010). BART: Bayesian Additive Regression Trees. *Annals of Applied Statistics 4*(1), 266–298.

Deshpande, S. K. (2024). **flexBART**: Flexible Bayesian regression trees with categorical predictors. *Journal of Computational and Graphical Statistics*.

Deshpande, S. K., R. Bai, C. Balocchi, J. E. Starling, and J. Weiss (2024, January). VCBART: Bayesian Trees for Varying Coefficients. *Bayesian Analysis -1*(-1).

Fisher, J. D., D. W. Puelz, and C. M. Carvalho (2020). Monotonic effects of characteristics on returns. *The Annals of Applied Statistics 14*(4), 1622 – 1650.

Fisher, J. D., D. W. Puelz, and S. K. Deshpande (2024). A Bayesian classification trees approach to treatment effect variation with noncompliance. arXiv:2408.00765.

Hahn, P. R. and C. M. Carvalho (2015). Decoupling shrinkage and selection in Bayesian linear models: A posterior summary perspective. *Journal of the American Statistical Association 110*(509), 435–448.

Hahn, P. R., C. M. Carvalho, D. Puelz, and J. He (2018, March). Regularization and Confounding in Linear Regression for Treatment Effect Estimation. *Bayesian Analysis 13*(1).

Hainmueller, J., J. Mummolo, and Y. Xu (2019, April). How Much Should We Trust Estimates from Multiplicative Interaction Models? Simple Tools to Improve Empirical Practice. *Political Analysis 27*(2), 163–192.

Hastie, T. and R. Tibshirani (1993). Varying-coefficient models. *Journal of the Royal Statistical Society Series B 55*(4), 757–796.

Hayfield, T. and J. S. Racine (2008). Nonparametric econometrics: The np package. *Journal of Statistical Software 27*(5).

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

Imbens, G. W. and D. B. Rubin (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.

Jackson, J. E. (1991). Estimation of models with variable coefficients. *Political Analysis 3*, 27–49.

Kalla, J. L. and D. E. Broockman (2023, January). Which Narrative Strategies Durably Reduce Prejudice? Evidence from Field and Survey Experiments Supporting the Efficacy of Perspective-Getting. *American Journal of Political Science 67*(1), 185–204.

Lee, Y. K., E. Mammen, and B. U. Park (2012). Flexible generalized varying coefficient regression models. *Annals of Statistics 40*(3), 1906 – 1933.

Li, Q. and J. S. Racine (2010). Smooth varying-coefficient estimation and inference for qualitative and quantitative data. *Econometric Theory 26*(6), 1607–1637.

Puelz, D., P. R. Hahn, and C. M. Carvalho (2017). Variable selection in seemingly unrelated regressions with random predictors. *Bayesian Analysis 12*(4), 969–989.

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

Rubin, D. B. (1980). Randomization analysis of experimental data: The Fisher randomization test. *Journal of the American Statistical Association 75*(371), 575–582.

Splawa-Neyman, J., D. M. Dabrowska, and T. P. Speed (1990). On the application of probability theory to agricultural experiments. Essay on principles. Section 9. *Statistical Science 5*(4), 465–472.

Sukhtankar, S., G. Kruks-Wisner, and A. Mangla (2022, July). Policing in patriarchy: An experimental evaluation of reforms to improve police responsiveness to women in India. *Science 377*(6602), 191–198.

Tibshirani, R. and J. Friedman (2020). A pliable lasso. *Journal of Computational and Graphical Statistics 29*(1), 215–225.