Subgroup Analysis: Pitfalls, Promise, and Honesty[*]

Marc Ratkovic[†]

January 10, 2020

Abstract

Experiments often focus on recovering an average effect of a treatment on an outcome. A *subgroup analysis* involves identifying subgroups of observations for which the treatment is particularly efficacious or deleterious. Since these subgroups are not preregistered but instead discovered from the data, significant inferential issues emerge. We discuss methods for conduct honest inference on subgroups, meaning generating valid p-values and confidence intervals which account for the fact that the subgroups were not specified *a priori*. Central to this approach is the split-sample strategy, where half the data is used to identify effects and the other half to test them. After an intuitive and formal discussion of these issues, we provide simulation evidence and two examples illustrating these concepts in practice.

Key Words: subgroup analysis, heterogeneous treatment effects, split sample, causal inference, machine learning

1 Introduction

In an experimental analysis, randomized assignment of a treatment variable allows for unbiased estimation of an average causal effect of the treatment. The average effects of interest are specified in advance by the researcher, and standard inferential tools allow estimation and testing of these effects. Every average effect, though, is itself a composite of lower-level

[*] We thank Jamie Druckman, Don Green, Adam Glynn, Zenobia Chan, and Stephanie Zonszein for feedback and comments.

[†] Assistant Professor, Department of Politics, Princeton University, Princeton NJ 08544. Phone: 608-658-9665, Email: ratkovic@princeton.edu, URL: http://scholar.princeton.edu/ratkovic

subgroup effects. A *subgrouping* is a partition of the sample into mutually exclusive subsets, normally split on observed covariates (e.g. Berry, 1990). A *subgroup* is one of these homogenous subsets, such as females, residents of a particular town, or white voters under thirty, and a *subgroup effect* is the average treatment effect for this subgroup. While randomization allows us to safely average over subgroups, a subgroup analysis turns this problem on its head: given data, how can we identify subgroups of the data where the treatment was most or least efficacious (e.g. Lagakos, 2006; Rothwell, 2005; Assmann et al., 2000)?

Subgroup effect estimation is crucial in at least four different settings. First, it can help identify the most impactful treatment from a set of possible treatments. In the face of increasingly complex designs, such as the conjoint analysis (Hainmueller, Hopkins and Yamamoto, 2014), the most effective treatment may involve a combination of two or three possible treatment conditions. Second, it can help characterize an ideal treatment for a given observation, which is of particular import when making policy prescriptions (Murphy, 2003). Third, a subgroup analysis can help with recent concerns over replicability. It may be that average estimate effects may fluctuate from one sample to the next, but this fluctuation may be attributable due to the fact that the samples have different distributions of underlying subgroups. Fourth, the analysis can help guide the researcher in designing the next experiment. As experimental analyses are part of a slow, careful accumulation ofcausal results (e.g. Samii, 2016), a subgroup analysis on a given experiment can help illuminate a likely mechanism and encourage future studies to focus on where an effect is most likely to be realized.

Despite this promise, the subgroup analysis raises two interrelated problems endemic to the design. A subgroup may be of theoretical interest and included in a preregistration plan prior to the experiment being conducted, and then tested as normal. When we say a *subgroup analysis*, though, we refer explicitly to an analysis where the goal is to identify effect heterogeneity among subgroups that were not specified prior to administering the experiment. The *ex post* nature of the analysis makes inference challenging. Simply reporting p-values from tests on subsets of the data that are not pre-registered is among the worst forms of data-dredging (e.g. Simmons, Nelson and Simonsohn, 2011). Second, the number of possible subgroups grows exponentially in the number of covariates. By the time we include a modest number of covariates, the number of potential subgroups can grow to the hundreds or thousands.

A wide range of off-the-shelf machine learning methods can be used to relax model heterogeneities in data (e.g. Montgomery and Olivella, 2018; Grimmer, Messing and Westwood, 2017; Green and Kern, 2012; Hill and Jones., 2014; LeBlanc and Kooperberg, 2010; Beck, King and Zeng, 2000). As we discuss below, these methods suffer in two regards. First, these off-the-shelf machine learning methods are tailored for optimal prediction, rather than optimal subgroup estimation. The difference between the two is that the best predictive values are driven by confounders, like partisanship or previous behavior, and most of the information in the data goes to learning these confounding effects rather than the treatment effects, which are often an order of magnitude smaller. Optimal subgroup estimation, instead, requires focusing on variables that drive the treatment effect, not those that drive the outcome. Second, these methods do not allow for valid inference on subgroup effects. Learning subgroup effects and conducting inference on them is too much to ask of any single dataset, regardless of the statistical method being used. Without reliable inferential claims, subgroup analysis devolves into fishing. We introduce a set of practices designed to allow for optimal estimation of subgroup effects as well as valid inferential claims to be made about these inferential effects.

In this work, we provide an overview of subgroup analysis aimed at the practical user. The discussion divides into three parts. The first details how a subgroup analysis fits into the design and pre-registration of an experiment. We show how to conduct an honest subgroup analysis, where the word honest takes on both technical and colloquial meanings. The key, as we discuss more below, is to think of a subgroup analysis as a process of discovery rather than a test, so the process itself needs to be specified in advance even if the outcome is unknown. We focus on how to generate optimal point estimates and valid uncertainty estimates, where the latter point has remained underdeveloped in the literature. We then revisit these concepts in a formal framework, highlighting how these ideas come into play when thinking about heterogeneity and inference. In this section, we discuss several methods that can be utilized for a subgroup analysis. The third section contains an illustrative simulation, showing how the split-sample approach reduces bias in subgroup estimation. The section also includes two worked-through examples, the first with a single binary treatment and the second illustrating how to conduct a subgroup analysis with multiple treatments and levels from a conjoint experiment. The conclusion discusses avenues of future work.

2 Design, Preregistration, and Honesty

The "replication crisis" that originated in several of our cognate fields has spurred a reconsideration of our experimental procedures (Gelman and Loken, 2014, provide a nice overview). At root, this crisis stems from the divergence between the theoretical guarantees of our means of inference and actual practice. Adjusting how we run experiments, including now-standard practice such as preregistration and a preanalysis plan, guards against the worst threats to the validity of our hypothesis testing. Our focus is on extending these same ideas to a subgroup analysis. We will describe a way for conducting honest subgroup analysis. We mean the word honest in two senses: first, formally, in that the procedures we discuss achieve a theoretically guaranteed error rate, but also second, in that an honest procedure creates statistical guardrails against misleading or deceptive inferences.

Our focus, then, is on not just estimating subgroup effects but inference on estimated or recovered effects. We first describe honesty in the context of testing an *ex ante* specified single null hypothesis, illustrating notions of validity, controlling an error rate, and honesty. In this section, we abstract what makes a testing procedure valid. We then move on to generalize these concepts to a subgroup analysis.

2.1 Honest and Efficient Inference on a Single Hypothesis

We focus on three separate concepts in thinking about inference on a single, pre-specified hypothesis: validity, honesty, and power. A test statistic for the hypothesis is valid if the false positive rate can be controlled by the researcher. A false positive occurs when a statistically significant result is observed even though the null hypothesis, normally of no causal effect, is true. We say that the false positive rate on a test is controlled at rate α if the researcher can guarantee that the proportion of statistically significant results that would be observed under the null is no more than α. Controlling the false positive rate is a first-order concern in experimental studies, as statistical significance serves as a crucial and necessary step in establishing that an estimated effect reflects a systematic relationship in the data. A testing procedure is *honest* if it results in a valid test statistic. For example, preregistering a design and hypothesis and then testing the hypothesis using a difference-in-means, as described in a preanalysis plan, results in an honest test.

An honest procedure has several components, and each can be violated if care is not exercised. That these violations lead to invalid p-values is well-understood (e.g. Wasserstein and Lazar, 2016; Gelman and Loken, 2014). Pre-registering handles these threats for the case of a single hypothesis, or small set of pre-declared hypotheses. First, the data generating process must be fixed, which, in practice, requires registering not just the design but how variables will be coded (Gelman and Loken, 2014). Second, the hypothesis must be generated independent of the data used to test them. Pre-specifying a hypothesis prior to running the experiment satisfies this requirement, though we discuss additional methods for doing so below in the context of a subgroup analysis. Third, the full set of hypotheses to be tested must be specified in advance. This guards against data-dredging, and again is satisfied by requiring the researcher to hypothesize effects prior to the experiment. Lastly, a valid test statistic must be used. If several exist, a more powerful method should be selected. In the case of a single hypothesis, using a t-test will achieve this goal, but the issue of power grows more important when learning hypotheses from the data.

2.2 Honest and Efficient Inference with a Subgroup Analysis

An experiment can help uncover three sets of causal effects: hypotheses specified in advance; hypotheses learned from the data; and discovered effects. The three classes differ on the persuasiveness of their evidence, ranging from highest to lowest. The first class is already addressed with current preregistration practices, so we move instead to the next two classes, which are the focus of subgroup analysis.

The second class are hypotheses learned and tested from the data. In order to return honest p-values, the procedure must maintain two of the attributes designed above. The procedure must not identify effects and test them on the same data. Recent work has advocated "sample-splitting" as a central feature of maintaining honesty (Wager and Athey, 2017; Athey and Imbens, 2016; Chernozhukov et al., 2018). In this framework, half the data is used to identify promising subgroups, and the other half of the data is used to test them. Were the same data to both generate hypotheses and test them, the p-values would not be valid; sample-splitting serves a crucial role in maintaining honesty (see, e.g. van der Vaart, 1998, ch. 25).

Concerns over power emerge when trying to estimate subgroup effects. While an average effect may be estimated off all the data, each subgroup is estimated over a smaller subset. Estimating relevant subgroups at this stage involves confronts an additional, and subtler, issue: the subgroup effects we are interested in have an impact an order of magnitude below the effect of the confounders. This has substantial implications on the estimation strategy. Most off-the-shelf machine learning methods try to predict the observed outcome as accurately as possible; this is a distinctly different concept than trying to estimate a causal effect. In many settings, the most important predictive variables are those that are the best known and least interesting. For intuition, consider the problem of predicting whether an individual exposed to a treatment condition votes. The strongest treatment effect may come from engaging in meaningful conversation with a canvasser, but this effect is an order of magnitude less than whether the respondent voted in the last election. A method tuned for prediction will spend quite a bit of power in the data learning the relationship between past and future voting, while a method sensitive to heterogeneity will ignore the past voting variable and focus primarily on variables that involve the treatment (Imai and Ratkovic, 2013). The predictive models are distracted by these known, but strong, confounders. Instead, the estimation strategy needs to target causal heterogeneities and avoid these known effects. Doing so involves rethinking standard estimation strategies, and we return to how to accomplish this below.

Inference on subgroup effects at this stage comes with two important caveats. First, since the data is split in half, so is the power of this method. This only seems fair, though, since we are asking two things of the data: identifying a subset of subgroups and then testing them. The second caveat is that both splits of the data come from the same, single experiment. The subgroups identified, then, may be the result of a peculiarity of this particular experiment, which serves as the key distinction between the first two classes of hypotheses.

A crucial question for the second class of hypotheses are how to identify them from the data. We discuss two different sets of methods below. The first point-identify subgroups in one split and test them in a second. The confidence intervals and p-values can be read off the test split. The second set return a fitted model of treatment effects, and must be explored by the researcher. Looking at plots of the estimated effects variable-by-variable is itself a form of exploration that must be accounted for; we recommend pre-registering the plots (say, all one-way or two-way effects across variables) and taking a Bonferroni-adjusted threshold for significance. For example, a researcher interested in learning subgroup effects across five variables, ignoring interactions among them, should use a split-sample machine learning method to estimate confidence intervals across each variable, but use a p-value threshold of 1/5 times their allowable false positive rate (say 0.1 or 0.05) on the subgroup effects.

The final set of subgroup effects are those for which we make inferential claims about the proportion of false positives in the entire set, rather than the false positive rate on any particular hypothesis. Rather than control the false positive rate, the probability that a given statistically significant effect is in-truth null, we control the false discovery rate: the proportion of discovered effects that are false. We describe a means of doing so below, but for this third class of effect, our inferential goals change. In the previous set of hypotheses, we are trying to make claims about how a treatment effect varies along a particular covariate. In this set, we are looking through a potentially massive set of hypotheses and attempting to discover plausible candidates to be tested in the next experiment.

This final set of effects should be considered when the number of potential hypotheses is massive. This may occur in two settings. First, if the researcher is considering all 3- or 4-way interactions, which is plausible in a conjoint setting, the number of possible subgroup effects may grow to the thousands or above quickly. Second, the researcher may have a large number of covariates. For example, if the pre-treatment covariates include a textual component, the term-document matrix may have thousands of unigrams or bigrams. In this case, there is little hope of testing each hypothesis, but we may be able to identify subgroups that may be worth testing in the next experiment.

2.3 Concrete Recommendations for Subgroup Analysis in a Pre-analysis Plan

For clarity, we provide our concrete recommendations to generate an honest subgroup analysis below:

1. Preregister the covariates, along with the level of interaction, that are going to define subgroups targeted for inference.

2. Preregister the method and, if applicable, the random seed with which it will be initialized. Utilize a method that estimates subgroup effects directly, to maximize power, and implements a sample splitting in this step, to maintain honesty of inference.

3. Preregister the marginal plots that will be used to look for heterogeneity. Every plot should be thought of as a degree of freedom, such that the p-value threshold should be Bonferroni adjusted. These uncovered effects can be discussed using the language of significance, but with the understanding that the process is conditional on the particular implementation of the experiment.

4. In order to define discovered effects, pre-register an acceptable false discovery rate and implement a method that achieves this rate. No inferential claims about statistical significance can be made about these, but interesting discoveries should be noted for future experimentation.

With a high-level discussion out of the way, we turn to a formal discussion of the problem and then discuss several options for implementing a subgroup analysis.

3 The Formal Framework

Treatment effects and subgroup effects are best characterized in the *potential outcomes* framework (Imbens and Rubin, 2015; Holland, 1986). We consider observations from a random sample $i \in \{1, 2, \ldots, N\}$. We observe for each observation an outcome Y_i, a vector and a vector of covariates X_i, with X an arbitrary covariate profile. We will denote as $z_i \in \{0, 1, 2, 3, \ldots, K\}$ the random vector taking value in one of $K + 1$ different treatment conditions, Z_i as its observed value, and Z as an arbitrary value that can be taken by the treatment. In the case of a single binary treatment, $z_i \in \{0, 1\}$. In a more complex setting, say with one treatment with three levels and another with four levels, $K = 11 = 3 \times 4 - 1$, the total number of treatment conditions beyond the control condition. The level $z_i = 0$ is reserved for the baseline level for each treatment. In this setting, each observation i has a potential outcome function $y_i(Z_i) = Y_i$ that maps each treatment level to an observed outcome.[1]

The average treatment effect for treatment condition $Z \in \{0, 1, 2, \ldots K\}$ is denoted as

$$\tau_i(Z) = y_i(Z) - y_i(0)$$

and with average effect

$$\tau(Z) = \mathbb{E}\left\{y_i(Z) - y_i(0)\right\}. \tag{1}$$

If the treatment is randomized, a difference-in-means estimate is unbiased for $\tau(Z)$.

We are also interested in the subgroup effect for observations with some covariate profile of interest X. We write this *conditional average treatment effect* or CATE as

$$\tau(Z; X) = \mathbb{E}\left\{y_i(Z) - y_i(0)|X_i = X\right\},$$

which is the treatment effect of condition Z for observations with covariate profile X.

We next turn to two central issues: how to estimate these heterogeneities effectively and how to conduct inference on them honestly.

3.1 Maximizing Power in Estimating Subgroup Effects

Uniformly powerful tests are simple when estimating average treatment effects, as t-tests or least squares models can returning unbiased and efficient estimates. Power grows more important in estimating subgroup effects. First, of course, each subgroup is characterized by a fraction of the data. Second, we need to differentiate between the best predictive model and

[1] We make the standard assumptions that each treatment has only one version, there is non-interference among units, and every treatment condition is realized with positive probability.

the model best suited to uncover subgroup effects. Off-the-shelf machine learning methods that find the best prediction focus on the most pronounced aspects of the data. This results in the method "learning" non-causal relationships in the data that are already known to, and often uninteresting to, the researcher. Efficient and powerful estimation of subgroup effects requires differentiating a predicted value from a subgroup effect, and directly targeting the latter.

To formalize, we follow Athey and Imbens (2016) and distinguish between two different types of estimation strategies. The first attempts tries to explain as much variance in the outcome as possible, given the treatment and outcomes. Taking $\mu(Z, X) = \mathbb{E}(Y_i|Z_i = Z, X_i = X)$, the first minimizes

$$\widehat{\mu}(Z, X) = \underset{\widetilde{\mu}}{\operatorname{argmin}}\, \mathbb{E}\left\{(Y_i - \widetilde{\mu}(Z_i, X_i))^2\right\}$$

which is the best predictive model. Off-the-shelf machine learning methods will attempt to minimize this predictive error.

The subgroup analysis can at its most general can be characterized as finding the combinations of potential treatments and covariate profiles that gives the largest value of treatment heterogeneity. The second set of methods attempt to explain as much treatment heterogeneity as possible,

$$\widehat{\tau}(Z, X) = \underset{\widetilde{\tau}}{\operatorname{argmin}}\, \mathbb{E}\left\{\left(\tau(Z_i; X_i) - \widetilde{\tau}_i(Z_i, X_i)\right)^2\right\}$$

We should favor methods that directly estimate $\widehat{\tau}$ rather than those that estimate $\widehat{\mu}$ and reconstruct $\widehat{\tau}$. The former group is efficient, while the latter are not. The difference between the two approaches is subtle, but important. The predictive approach will attempt to learn the best model of the outcome, which includes more than the treatment effects of interest. Consider partitioning the predictive loss function into two components: a component that varies with the treatment and one that does not. The second element, that does not vary with the treatment, is of no interest to the researcher, and any information in the data spent on this component is wasted. The treatment heterogeneity approach only considers variance in the outcome that can be explained with the treatment variable. The predictors that are of no interest are differenced-out , focusing estimation on covariates that drive the treatment effect.

Characterized in this way, two questions present themselves: how can we discover these heterogeneities and how can we conduct inference on our findings? We turn to each question in turn.

3.2 Estimating Treatment Effect Heterogeneity

We discuss two approaches that can be used to generate efficient estimates of treatment effects. The first approach was most recently advanced in a series of papers by Susan Athey and colleagues (Athey and Imbens, 2016; Wager and Athey, 2017; Athey, Tibshirani and Stefan Wager. Annals of Statistics, 2019). The loss function

$$\widehat{\tau}(Z,X) = \underset{\widetilde{\tau}}{\operatorname{argmin}}\, \mathbb{E}\left\{ \left(\tau(Z_i;X_i) - \widetilde{\tau}(Z_i,X_i)\right)^2 \right\}$$

is not feasible, since we do not know the true function $\tau(Z;X)$. The authors show that the estimate

$$\widehat{\tau}(Z,X) = \underset{\widetilde{\tau}}{\operatorname{argmin}} -\mathbb{E}\left\{ \widetilde{\tau}(Z_i,X_i)^2 \right\}$$

will optimally recover treatment effect heterogeneity.[2] The loss function offers a nice practical interpretation: a method that attempts to explain as much treatment effect heterogeneity as possible is optimal for recovering subgroup effects. The authors have then developed a suite of tree- and forest-based methods that efficiently estimate subgroup effects.

A second approach for avoiding confounders involves removing their effect prior to identifying the subgroups. Recently advocated by Chernozhukov et al. (2018), the approach requires a two-step procedure. In the first step, the effect of the confounding variables is taken out of the outcome and treatment variable, generating variables

$$\widetilde{Y}_i = Y_i - \mathbb{E}(Y_i|X_i); \quad \widetilde{Z}_i = Z_i - \mathbb{E}(Z_i|X_i).$$

[2] The derivation relies crucially on having an unbiased estimate of $\widehat{\tau}(Z_i,X_i)$, which itself requires a split-sample approach to estimation. We return to this point below, but see Athey and Imbens (2016) for a derivation.

We refer to these as the "partialed out" variables (Robinson, 1988; Chernozhukov et al., 2018; Neyman, 1979), since we have subtracted off the impact of the confounders. Importantly, any machine learning method can be used to conduct the partialing out; see Chernozhukov et al. (2018) for formal details. We then run a method for discovering subgroups on these partialed out values.

Importantly, if the treatment is properly randomized by the experimenter, such that the value of $\mathbb{E}(Z_i|X_i)$, this value need not be estimated. In this case, a predictive model and a model incorporating inverse probability of treatment weights are asymptotically indistinguishable. If, on the other hand, there is some fear that that the experiment was not perfectly executed, then the researcher may prefer to utilize some method to adjust for any bias that can be accounted for by the covariates.

Both methods were designed not just with an eye to estimation but also inference. They are both implemented using a split-sample approach in order to ensure honest inference, a point to which we turn next.

3.3 Split-Sample Approaches for Honest Inference

As discussed above, the goal with a subgroup analysis is not just to identify relevant subgroups but to inferential claims about those that are uncovered. The classes of methods described directly above were designed to be implemented using a split sample. A split sample involves taking the data and simply splitting it into two equally sized subsets; the splits may be done completely at random or may be done with respect to any blocking in the experiment.

For tree-based methods, one split of the data is used to learn the tree structure, and then the second split of the data is used to conduct inference at each terminal node; a set of these trees may be aggregated up to a forest. Similarly, using the partialing-out approach, one split of the data is used to remove the effect of the confounders and the second is used to learn subgroup effects.

Previous work has used the split-sample approach to recover honest estimates of regression coefficients (Robinson, 1988; Chernozhukov et al., 2018) or average treatment effects (Wager and Athey, 2017). We extend this approach to show, in some generality, that the split-sample approach can be used to guarantee valid inference on a test statistic. Specifically, we describe an honest procedure where the split-sample approach can be used to conduct valid inference. The method involves a split-sample approach. The first split of the data is used to select a subgroup effect to be tested, and a null hypothesis of zero effect is assumed. The second split of the data is used to test this hypothesis, using any valid test. We show that this method is honest.

To formalize, we assume a set of null hypotheses, each corresponding with a potential subgroup effect upon which we may wish to conduct inference. We will denote null hypothesis h as $\mathcal{H}_0$ with $h \in \{1, 2, \ldots, H\}$. The observed sample is denoted $\mathcal{S}$. We assume a test statistic $\widehat{t}_h$ for null $\mathcal{H}_0^h$ and a significance threshold as t_h^* such that the researcher can control the false positive rate on the test. For example, $\widehat{t}_h$ may be a z-statistic and t_h^* the familiar threshold of 1.96. An effect is significant if the test-statistic is larger than the threshold. We also assume that $\widehat{t}_h$ is estimated and t_h^* selected to give the same false positive rate across all hypotheses. We use the data to learn a promising subgroup, say $h(\mathcal{S})$, and make a null hypothesis about it, $\mathcal{H}_0^{h(\mathcal{S})}$, to differentiate the null from a hypothesis made independent of the data, $\mathcal{H}_0^h$. Under null hypothesis $\mathcal{H}_0^h$, a false positive occurs if $\mathbb{1}(|\widehat{t}_h| > t_h^*)$ and the false positive rate we wish to achieve is

$$\mathbb{E}\left\{\mathbb{1}\left(|\widehat{t}_h| > t_h^*\right)|\mathcal{H}_0^h\right\} = \mathbb{E}\left\{\mathbb{E}\left(\mathbb{1}(|\widehat{t}_h| > t_h^*)|\mathcal{S}, \mathcal{H}_0^h\right)\right\},$$

where the outer expectation is over repeated samples, under the null. The issue in a full-sample subgroup analysis is that we have consulted the data to learn the hypothesis $\mathcal{H}_0^{h(\mathcal{S})}$, which renders our false positive rate incorrect:

$$\mathbb{E}\left\{\mathbb{E}\left(\mathbb{1}(|\widehat{t}_h| > t_h^*)|\mathcal{S}, \mathcal{H}_0^h\right)\right\} \neq \mathbb{E}\left\{\mathbb{E}\left(\mathbb{1}(|\widehat{t}_h| > t_h^*)|\mathcal{S}, \mathcal{H}_0^{h(\mathcal{S})}\right)\right\} \tag{2}$$

Intuitively, if we use the data to select a promising subgroup effect, then it is biased towards being significant, since we are using the same data to test the hypothesis that we used to uncover it.

Instead, assume we have two equally-sized splits of the data, $\mathcal{S}_1$ and $\mathcal{S}_2$, where we use $\mathcal{S}_1$

to learn a promising subgroup and $\mathcal{S}_2$ to test it, our false positive rate is

$$\mathbb{E}\left\{\mathbb{E}\left(\mathbb{1}(|\widehat{z}| > z^*)|\mathcal{S}_2, \mathcal{H}_0^{h(\mathcal{S}_1)}\right)\right\}$$

By decoupling selection and testing of the hypothesis, this procedure returns a valid test,

$$\mathbb{E}\left\{\mathbb{E}\left(\mathbb{1}(|\widehat{z}| > z^*)|\mathcal{S}_2, \mathcal{H}_0^{h(\mathcal{S}_1)}\right)\right\} = \mathbb{E}\left\{\mathbb{1}(|\widehat{t}_h| > t_h^*)|\mathcal{H}_0^h\right\}.$$

To see this, first fix $\mathcal{S}_1$, which fixes $\mathcal{H}_0^{h(\mathcal{S}_1)}$. In this setting, any false positive is attributable to variance in $\mathcal{S}_2$, which returns a valid test. Since we can imagine, then, sampling over $\mathcal{S}_1$ repeatedly, the test is valid since any connection between selecting the hypothesis and testing it is broken.

In practice, we have shown that a valid test of a null hypothesis generated from a split sample will achieve the nominal error rate. There are, of course, several practical issues that may interfere with this result. First, there may be something about the particular experiment that is not representative of the full population. Irregularities in administering the experiment will bias any estimate and invalidate the procedure. Second, this method falls prey to issues of multiple testing as well as any. The preferred approach, which we recommend, is to pre-specify how many tests are anticipated and then Bonferroni-correct the p-values.

3.3.1 Adjusting for Multiple Hypotheses

The false positive rate is the probability of a false positive on a single hypothesis, over repeated samples. The family-wise error rate is the probability of achieving a false positive amongst a family of K hypotheses. Define the number of statistically significant effects from a set of K hypotheses $h \in \mathcal{H}^K$ as

$$V = \sum_{h \in \mathcal{H}^K} \mathbb{1}(|\widehat{t}_h| > t_h^*).$$

Denote $\mathcal{H}_0^K$ as the composite null hypothesis that every hypothesis in $\mathcal{H}_K$ is true. The family-wise error rate is

$$\mathbb{E}\left\{V > 0|\mathcal{H}_0^K\right\}$$

Of course, as we test more hypothesis, the probability of a false positive increases. The simplest way to adjust for this issue is with a Bonferroni correction. The method is a simple, but valid, method for adjusting for multiple-hypotheses. The correction simply involves replacing a p-value threshold of p^* with p^*/K, and doing so allows control of the familywise error rate (e.g. Esarey and Summer, 2015).

3.4 Controlling the False-Discovery Rate

The previous discussion has focused on testing a particular hypothesis, whether it is specified in advance or learned from the data. We turn now from a hypothesis-testing framework to a hypothesis-discovery framework. The goal shifts from rejecting a particular null to estimating a set of possible effects, but controlling the proportion of effects that are false.

Define as R the number of selected effects. The false discovery rate is

$$\mathbb{E}\left\{V/R\right\} = \mathbb{E}\left\{V/R|R > 0\right\} \Pr(R > 0)$$

where the second form of the expression excludes the case where $R = 0$. The standard method for controlling the false discovery rate is the Benjamini-Hochberg Procedure. The method works in the following fashion. Given a large set of hypothesis and desired false discovery rate α, the Benjamini-Hochberg Procedure works in two steps. First, the K estimated p-values on each test are ordered from smallest to largest, as $\{p_{(1)}, p_{(2)}, \ldots, p_{(K)}\}$. Then, the selected set are the tests associated with p-values such that

$$\left\{k : \; p_{(k)} < \frac{k}{K}\alpha\right\} \tag{3}$$

This procedure guarantees a false discovery rate on the set of discovered hypotheses of below α. The acceptable false discovery rate should be specified in advance; 0.1 and 0.05 are standard values.

The discovered effects should be considered as a group, as we cannot make claims about any one hypothesis. The goal here is to reduce thousands of possible effects to a manageable number that can be explored in the next round of experimentation. The full set of hypotheses and acceptable false discovery rate should be specified in advance.

3.5 Estimation Advice

In this section, we provide advice on particular algorithms that can be used for subgroup analysis. As these methods are commonly updated, we expect that our concrete recommendations may be out of date at some point. Therefore, we strive here and above to highlight our reasoning as well as our recommendations.

If the goal is using a model-based machine learning method in order to learn subgroup effects, then the researcher should not implement methods that do not engage in sample-splitting or, at the least, make some differentiation between estimating the treatment effect and prediction. Examples include off-the shelf predictive tools like random forests or their Bayesian variants (e.g. Hill and Jones., 2014) and super learners that average over multiple methods (e.g., Grimmer, Messing and Westwood, 2017). While these are excellent for prediction, they are not optimized for uncovering subgroup effects or properly tuned for inference.

If the experiment has a single, binary treatment, we recommend implementing the generalized random forest algorithm of Athey, Tibshirani and Stefan Wager. Annals of Statistics (2019). The method uses a split-sample approach, with one sample to learn a tree structure then the other for inference. This process is repeated and embedded within a random forest, such that all the data is used at some point to either learn a heterogeneous tree structure or to estimate a treatment effect. We implement this method in an application below.

The case of multiple treatments does not lend itself as well to the tree and forest approach, since any number of heterogeneous effects now exist and may need estimation. Here, we suggest using some version of a sparse model combined with a split-sample approach in order to find an effect. The method would split the data, and model the outcome in terms of the covariates on one half of the data. A large set of controls, interactions, and higher-order terms should be included. In the second, the partialed out outcome should be regressed on a large set of covariates comprised of *treatment* × *covariate* interactions. We recommend a sparse model, such that it takes a large number of these interactions and returns some subset of the most relevant. The standard LASSO (e.g. Hastie, Tibshirani and Friedman., 2013) does not return standard errors, so we offer two alternatives. First, recent work uses the LASSO to select covariates, but then runs least squares on a subset of the selected covariates. The confidence intervals and p-values on these uncovered effects are valid (Belloni et al., 2017). A second set of methods use a Bayesian framework to recover a subset of relevant effects, and either (Park and Casella, 2008; Carvalho, Polson and Scott, 2010, e.g.); see Ratkovic and Tingley (2017) for a full discussion and extensions.

If you have a large number of subgroups or treatment-covariate interactions, such as tens of thousands or more, we recommend the Benjamini-Hochberg approach. We have seen recent interest in the method (e.g. White et al., 2018), and encourage its widespread adoption in situations where a potentially vast number of subgroups exist.

4 Simulation Evidence and Applied Example

We next turn to illustrate a few of the concepts discussed above. In a simulation setting, we compare two machine learning methods that have been used for identifying effect heterogeneity: Bayesian Additive Regression Trees (*BART* Hill, Weiss and Zhai, 2011), which do not implement a split-sample and partialing-out approach, and Causal Forests (Athey, Tibshirani and Stefan Wager. Annals of Statistics, 2019) , which do. We find the latter returns subgroup estimates notably less biased and confidence intervals that are more reliable.

We then include two applications, to an audit study (see Butler and Crabtree's chapter in this volume) and a conjoint analysis (See Bansak, et al.,'s chapter in this volume).

4.1 Simulation Evidence

We illustrate the basic insight, that the split-sample approach combined with partialing out, can reduce bias and lead to more reliable inference. For this simulation, we generate a set of three covariates, each independent and identically standard normal, denoted $\{X_{i1}, X_{i2}, X_{i3}\}$. Denote $S_i = \textbf{sign}(X_{i1} + X_{i2})$, so it takes a value of $+1$ if the sum is positive and -1 if the sum is negative. We use two different simulation settings that vary with how we generate the binary treatment variable Z_i. In the first, we generate the treatment as a coin flip; in the second, the treatment probability is a function of S_i:

$$\begin{aligned} &\text{Setting 1: } Z_i|S_i \sim \text{Bern}(\pi_i); \pi_i = 0.5 \\ &\text{Setting 2: } Z_i|S_i \sim \text{Bern}(\pi_i); \pi_i = \frac{1}{1+\exp(-S_i)} \end{aligned}$$

In each setting, we generate the outcome as

$$Y_i = S_i + \epsilon_i$$

where ϵ_i is itself standard normal and independent of the covariates. Note, importantly, that the causal effect for each observation is zero: Z_i does not enter the outcome. By this means, any effect we find as significant is a false positive. We compare the performance of these methods on estimation and inference on six different subgroups, given by

$$\{X_{i1} > 0, X_{i1} \leq 0, X_{i2} > 0, X_{i2} \leq 0, X_{i3} > 0, X_{i3} \leq 0\}.$$

We implement two methods: the causal forest (CF) of Athey, Tibshirani and Stefan Wager. Annals of Statistics (2019) and Bayesian Additive Regression Trees ($BART$), which have been used in the past for subgroup analysis and causal estimation (e.g. Hill, Weiss and Zhai, 2011; Green and Kern, 2012, with and without sample-splitting, respectively). The methods are similar in that they both rely on a collection of trees to model the outcome (see, e.g. Montgomery and Olivella, 2018; Hill and Jones., 2014), and are designed to identify discontinuities in the data like those induced by S_i. The methods differ fundamentally in how they model the treatment effect: CF utilizes a split-sample strategy while $BART$ uses the full sample to model the outcome given the treatment and covariates. To recover the treatment effect from $BART$, we have the method predict the outcome by the treatment and covariates. We then estimate the treatment effect using these values, and use the posterior to generate uncertainty.

The intuitive arguments given above suggest that $BART$ will exhibit more bias than CF

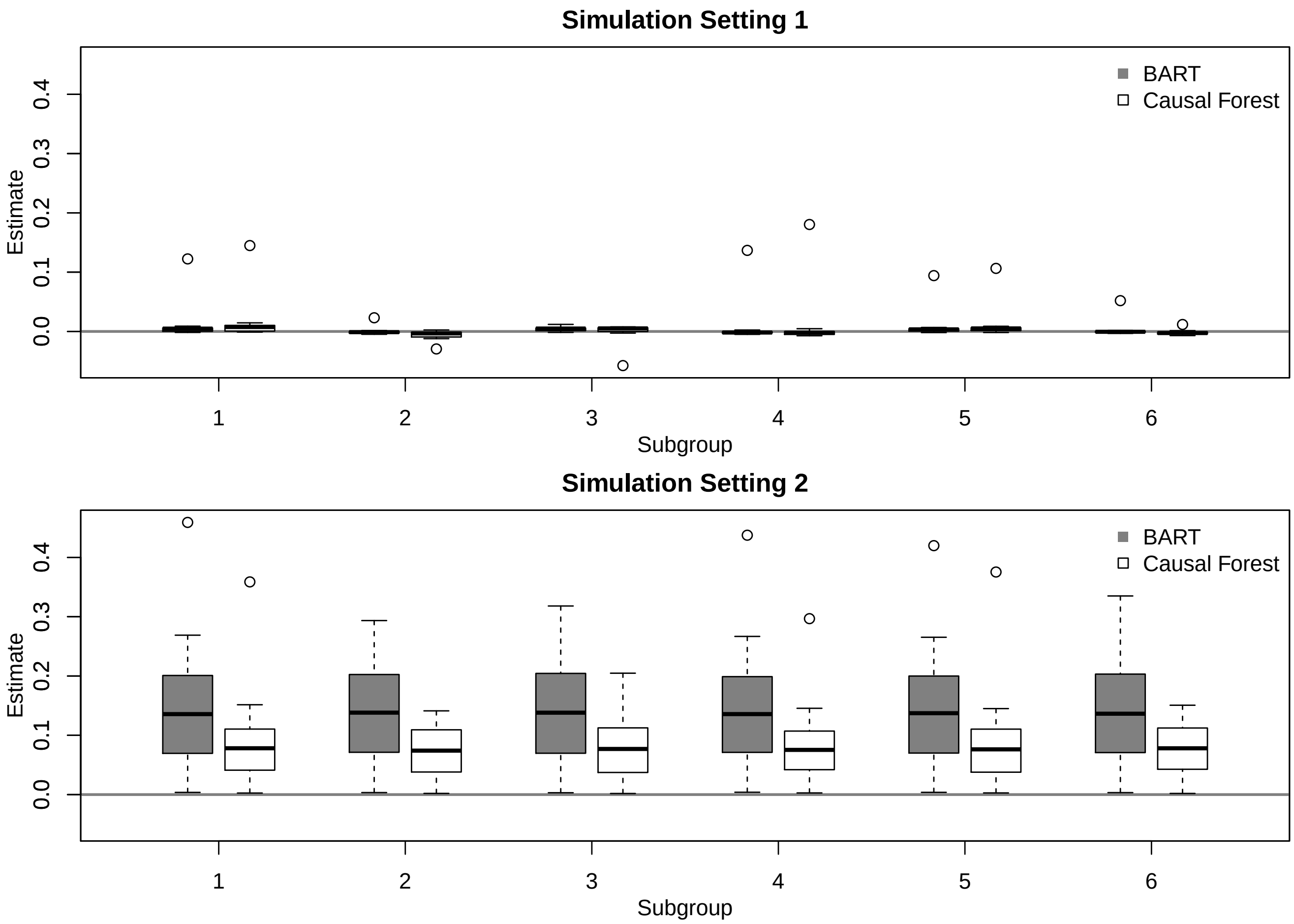

Figure 1: **Subgroup Estimates.** Estimated treatment effect by subgroup for BART (gray) and causal forests (white). Both methods are an average of trees, and there are only three variables, so both should perform well on this data. We see in Simulation 2 that causal forests have only about half the bias of BART.

in this second setting for two reasons: first, it does not model the treatment as a function of the covariates and it uses the same data to both model the outcome and generate estimates and uncertainty intervals. CF, on the other hand, splits the data in half, uses half to learn a forest structure to model the treatment effect, and the other to generate point and uncertainty estimates. This second procedure should generate estimates with less bias and better coverage, which we do indeed see.

We start with results on point estimation, presented in Figure 1. The top figure contains the results from Simulation 1, where the treatment is a coin flip, and the bottom for where the treatment probability varies with S_i. Both methods perform well when the treatment

		Subgroup					
		1	2	3	4	5	6
Setting 1	BART	0.94	0.93	0.91	0.92	0.95	0.92
	Causal Forest	0.94	0.87	0.89	0.90	0.93	0.92
Setting 2	BART	0.11	0.13	0.12	0.12	0.10	0.09
	Causal Forest	0.59	0.63	0.61	0.61	0.63	0.55

Table 1: **Coverage of 90% Uncertainty Intervals, by Subgroup.** The first two rows contain coverage results for the first simulation setting, the bottom two rows for the second setting. We see that both methods achieve nominal coverage when the treatment probability is homogenous. In the second setting, we see deterioration by both methods. The 90% confidence intervals for causal forests still cover the truth about 60% of the time, while the 90% posterior interval only contains the truth about 10% of the time.

probability is constant across units. In the bottom figure, giving the results from the second setting, we see that neither method is unbiased across subgroups. Causal forests, though, have about half the bias of BART. Note that we are not advocating exclusively for the causal forest algorithm per se, as algorithms and methods are constantly improving, but arguing that the split-sample approach should be preferred in a subgroup analysis, as it leads to more reliable inference.

Bias is not the whole story, though. We are focused on inference, asking whether we can trust our confidence intervals from a subgroup analysis. The results are shown in Table 1. The first two rows contain coverage results for the first simulation setting, the bottom two rows for the second setting. We see that both methods achieve nominal coverage when the treatment probability is homogenous. In the second setting, we see deterioration by both methods. The 90% confidence intervals for causal forests still cover the truth about 60% of the time, while the 90% posterior interval only contains the truth about 10% of the time. Put differently, in this setting where there is no treatment effect, a 90% credible interval from $BART$ that does not engage in a split-sample approach will make a false positive about 90% of the time. For causal forests, using a confidence interval derived by a split-sample approach, the false positive rate drops to about 40%–not perfect, but notably better.

4.2 Applied Example: A Field Experiment with a Single, Binary Treatment

In a recent audit study, Butler and Broockman (2011, see also Kalla, et al. in this volume) emailed US state legislators, varying whether the e-mail is sent from a constituent with a stereotypically black name (*DeShawn*) or white name (*Jake*); see the original paper for a full description of the design. The outcome we consider is whether the email receives a reply, where we take *DeShawn* as the treatment condition and *Jake* as the control condition. We consider subgroups given by party of the legislator (Republican, Democrat) and the legislator's race (white, black, Hispanic). We omit any subgroups that do not include at least 100 respondents, leaving us with the five main effects (Republican, Democrat, white, black, Hispanic) and four interactive effects (white Republicans, white Democrats, black Democrats, and Hispanic Democrats) for a total of 9 overlapping subgroups.[3] In order to adjust for making 9 comparisons, we use a Bonferroni correction. so we lower our p value from 0.1 to $p^* = 0.1/9 = 0.011$, which raises our critical value on the z-statistic on the difference in means from 1.64 to 2.54. The key attributes of this process for pre-registering is listing the number of subgroups, the threshold for dropping subgroups due to sample size, and that a Bonferroni correction will be used.

We estimated the causal effect for each subgroup using a causal forest. The results by subgroup can be found in Table 2. After the Bonferroni correction, we find three statistically significant effects.

First, we find that Republicans are less likely to return an email from the *DeShawn* condition. We also find a practically identical effect for white Republicans, largely because 97.4% of Republican legislators are white. While Republicans are less likely to respond to DeShawn, we find no significant main effect for Democrats. We find, though, that black Democratic legislators are more likely to respond to DeShawn than to Jake. Our results correspond with what was found in the original study, with two exceptions. First, we find that our Bonferroni correction eliminates one result that was found marginally significant ($p = 0.07$) in the original study, that white Democratic legislators are more likely to respond

[3] Since the vast majority of Republicans are white, dropping small subgroups removes black and Hispanic Republicans from our analysi.

	Estimate	SE	z	n
Main Effects				
Republican	-0.06	0.02	-2.80	2170
Democrat	0.02	0.02	0.80	2689
Black	0.12	0.05	2.47	349
Latino	0.06	0.09	0.69	141
White	-0.04	0.02	-2.33	4269
Interaction Effects: Republican ×				
White	-0.06	0.02	-2.85	2114
Interaction Effects: Democrat ×				
Black	0.13	0.05	2.60	343
Latino	0.06	0.10	0.60	115
White	-0.01	0.02	-0.46	2155

Table 2: **Subgroup Effects from Audit Experiment.** Estimated effects by subgroup for the Butler and Broockman study. All subgroups with under 100 people were omitted. The Bonferonni-corrected critical value for the z statistic is 2.54.

to Jake. Our findings on white Republicans are the same as those reported in the original study, and we find the effect attributed to minority Democrats (blacks and Hispanics pooled) to be driven by black legislators. The central methodological argument, though, is that p-values should be adjusted when conducting multiple tests on subgroups.

4.3 Applied Example: A Conjoint Experiment with Multiple Treatment Conditions

We next reanalyze data from a conjoint experiment. The original analysis considered the effect of varying dimensions of international climate agreement on respondent preferences given over three countries (UK, US, DE); see Bechtel and Scheve (2013) for a complete discussion of the design. Proposals were varied by the expected costs, how many countries participated, level of sanctions levied against violators, level of cuts required, the organization monitoring compliance, and whether costs would be distributed proportional to historical or current pollution rates. Moderators collected on respondents include gender, ideology, age, country, and whether the respondent is likely to engage in reciprocity, as measured in a two-player public goods game after the experiment.

We conduct our subgroup analysis in two stages: we use half the respondents to learn interesting subgroups and the other half to test them. For the sake of preregistration, we would characterize the set of subgroups we are considering, register the seed we are using to split respondents, the statistical method we are using to learn them in the first split, and that we are Bonferroni correcting for the number selected in the second split. In particular, we consider all *treatment* × *moderator* combinations, resulting in 368 possible subgroups being assessed. We then separate the observations in two equal splits. On the first split, we use the LASSO model of Belloni et al. (2017), which returns a subset of subgroups with estimated non-zero effect (for an overview of variable selection methods, see Ratkovic and Tingley (2017)). We then enter the selected subgroup covariates into a regression model on the second split, returning standard errors clustered by respondent and Bonferroni-adjusting the p-values for the number of selected covariates.

When constructing the covariates to measure a subgroup effect, we generate covariates that capture the causal effect of a given variable in a given subgroup (see, e.g., Bansak, 2018, for more). The covariates are constructed such that, within each subgroup, the treatment group and control group are contrasted; outside the subgroup, the covariate is set to zero. Specifically, let subgroups be denoted by indicator variables $g_{is}, s \in \{1, 2, \ldots, S\}$ where

$$g_{is} = \begin{cases} 1; & \text{observation } i \text{ in subgroup } s \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

Consider indicator variable for observation i, denoted $Z_{ik}, k \in \{1, 2, \ldots, K\}$. Denote as $\overline{Z}_{ks}$ the mean number of observations in subgroup s receiving treatment k. Then, our covariate for estimating the effect of treatment k on subgroup s is then

$$x_{isk} = \begin{cases} Z_{ik} - \overline{Z}_{ks}; & Z_{is} = 1 \\ 0; & g_{is} = 0 \end{cases} \tag{5}$$

This covariate is constructed such that the coefficient from regressing the outcome on this covariate gives the difference in means in that subgroup. It does so by zeroing out observations out of the subgroup, but creating a contrast between the treated and control in that subgroup. Note that this covariate could be constructed instead to contrast those in treatment condition k to some other condition. This covariate treats all the treatment levels except for k as the baseline, so coefficients on this covariate should be interpreted as the mean difference between the given treatment and the average of all other treatment levels for this variable. For example, we find a coefficient on *United.States* × *Cost: $267* of 0.029. This should be interpreted as, among respondents in the US (the subgroup), treaties that cost $267 per capita were less favored than the average across other possible levels by 2.9 percentage points.

Results from this procedure can be found in Table 3. The left column contains the results from a regression on the selected covariates on the first split of the data; the right column contains the same results from the second split. The righthand regression can be used for unbiased estimation and valid inference, since it was estimated off different data than that used to estimate the subgroups. The table reports point-estimates, standard errors clustered by respondent, and significance based off Bonferroni-corrected p-values. We do not report significance on the main effects, as captured by coefficients without interactions, since those are presumed to have been pre-registered and tested as normal; the main effects are included only as controls.

The discovered effects comport with intuition. In terms of main effects, we find more expensive plans to be less popular. Similarly, only having rich countries pay, having fewer participants, and higher sanctions are less preferred. In terms of subgroup effects, we use the second column to assess statistical significance, but focus most on those estimates which are stable and significant in both splits. Conservatives disfavor allowing Greenpeace to model compliance, and those with a high reciprocity score do not want the lowest number of participating countries (20). Respondents low on the environmentalist measure want no sanctions, while those scoring higher on the environmentalist measure disfavor fewer countries participating in the treaty, lower cuts, and their government monitoring the outcomes. The authors included a series of ad-hoc, interactive models in the supplemental materials of the original study, suggesting that they were interested in uncovering subgroup effects. Our primary argument here is that uncovering these effects should be done in a systematic fashion, in this setting using a machine learning method on half the data and a regression on uncovered effects in the other. Doing so helps to ensure valid inference on subgroup effects.

5 Conclusion

The subgroup analysis has remained a difficult part of experimental analysis, specifically because it can too-easily brush up against data dredging and p-hacking. At the same time, potential insights in experimental data should not be left unreported simply because they were not specified in advance.

Instead, our focus has been on reviewing and presenting a set of statistically rigorous concepts and methods with which we can conduct inference on subgroup effects. The methods, as we discuss, offer a higher bar of proof than a pre-specified hypothesis. The split-sample approach cuts the data we are testing the hypothesis on in half, where we use half the data to find promising hypotheses and the other half to test them. This leads to a decrease in power, but this decrease is the necessary consequent of not pre-specifying and pre-registering the hypotheses. Similarly, the false discovery rate may find promising patterns in the data, but the approach does not allow for inference on any particular hypothesis. These costs, though, allow for statistical guarantees that can help convey nuance in experimental data that would otherwise be ignored.

Looking forward, several open questions remain. First, methods for conducting subgroup analysis in more complex designs, like instrumental variables, mediation, and regression discontinuity designs are still open questions. Similarly, experiments with time-varying treatments and continuous treatments raise even more complex issues. In this chapter, we tried to look ahead past the current methods of the day for underlying principles: regardless of how a subgroup analysis is performed, and on what design, the split-sample approach and partialing out covariates will serve as a common thread through these analyses.

	Dependent variable: Support of Climate Change Treaty	
	(1)	(2)
Main Effects		
Cost: \$53	0.118 (0.015)	0.127 (0.015)
Cost: \$ 107	0.079 (0.009)	0.066 (0.008)
Cost: \$ 213	−0.092 (0.009)	−0.100 (0.008)
Cost: \$267	−0.096 (0.015)	−0.120 (0.015)
Only rich countries pay	−0.051 (0.008)	−0.049 (0.007)
20 of 192 participants	−0.010 (0.012)	−0.021 (0.012)
160 of 192 participants	0.020 (0.014)	0.028 (0.014)
Sanctions: \$11	0.026 (0.009)	0.044 (0.009)
Sanctions: \$43	−0.022 (0.008)	−0.025 (0.008)
Indep. Commission Monitors	0.034 (0.007)	0.037 (0.007)
Interaction Effects		
Female × Cost: \$53	0.029 (0.014)	0.050^{***} (0.014)
Female × Cost: \$267	-0.047^{**} (0.014)	−0.026 (0.013)
Female × 80% of emissions cut	0.018 (0.009)	0.004 (0.010)
Female × Sanctions: \$11	0.029 (0.012)	0.016 (0.012)
Conservative × Greenpeace Monitors	-0.079^{***} (0.012)	-0.074^{***} (0.011)
Liberal × 160 of 192 participants	0.020 (0.013)	0.004 (0.012)
Liberal × Your government monitors	−0.015 (0.011)	−0.001 (0.011)
Reciprocity: high × Cost: \$53	0.021 (0.014)	0.002 (0.014)
Reciprocity: high × Cost: \$267	−0.020 (0.014)	−0.010 (0.013)
Reciprocity: high × 160 of 192 participants	0.012 (0.013)	0.021 (0.013)
Reciprocity: high × 20 of 192 participants	-0.056^{***} (0.013)	-0.040^{**} (0.013)
Reciprocity: high × 80% of emissions cut	0.020 (0.010)	0.033^{**} (0.010)
Env: low × Cost: \$53	0.039 (0.014)	0.050^{***} (0.014)
Env: low × Sanctions: None	0.051^{***} (0.011)	0.047^{***} (0.011)
Env: low × Sanctions: \$43	−0.017 (0.013)	-0.046^{***} (0.013)
Env: high × 160 of 192 participants	0.043^{**} (0.014)	0.036 (0.014)
Env: high × 20 of 192 participants	-0.078^{***} (0.013)	-0.084^{***} (0.014)
Env: high × 80% of emissions cut	0.005 (0.010)	0.009 (0.010)
Env: high × 40% of emissions cut	-0.041^{***} (0.008)	-0.042^{***} (0.008)
Env: high × Your government monitors	-0.038^{**} (0.012)	-0.045^{***} (0.012)
United.Kingdom × Cost: \$53	0.037 (0.015)	0.005 (0.016)
United.Kingdom × Cost: \$267	-0.071^{***} (0.017)	−0.042 (0.017)
United.States × Cost: \$267	-0.064^{***} (0.015)	−0.029 (0.016)
United.States × Only rich countries pay	-0.041^{*} (0.014)	−0.033 (0.014)
Note:		*p<0.1; **p<0.05; ***p<0.01; *p*-values Bonferroni corrected

Table 3: **Split-Sample Estimates from Conjoint Analysis.** Results from the split-sample subgroup analysis from the conjoint experiment in Bechtel and Scheve (2013), with standard errors in parantheses. In the first split of the data, the covariates above were selected using a LASSO. These covariates were then entered into a linear model using the second split. Results are from the linear model in the first split (left) and second (right). Standard errors clustered by respondent.

.

References

Assmann, Susan F., Stuart J. Pocock, Laura E. Enos and Linda E. Kasten. 2000. "Subgroup Analysis and Other (mis)uses of Baseline Data in Clinical Trials." *The Lancet* 355(9209):1064–1069.

Athey, Susan and Guido Imbens. 2016. "Recursive partitioning for heterogeneous causal effects." *Proceedings of the National Academy of Sciences of the United States of America* 113(27):7353–7360.

Athey, Susan, Julie Tibshirani and software] Stefan Wager. Annals of Statistics, forthcoming. [arxiv. 2019. "Generalized Random Forests." *Annals of Statistics* . Forthcoming.

Bansak, Kirk. 2018. "A Generalized Framework for the Estimation of Causal Moderation-Eects with Randomized Treatments and Non-RandomizedModerators." Working Paper.

Bechtel, Michael M and Kenneth F Scheve. 2013. "Mass support for global climate agreements depends on institutional design." *Proceedings of the National Academy of Sciences* 110(34):13763–13768.

Beck, Nathaniel, Gary King and Langche Zeng. 2000. "Improving Quantitative Studies of International Conflict: A Conjecture." *American Political Science Review* 94(1):21–35.

Belloni, Alexandre, Victor Chernozhukov, Ivan Fernandez-Val and Christian Hansen. 2017. "Program Evaluation and Causal Inference With High-Dimensional Data." *Econometrica* 85(1):233–298.

Berry, Donald. 1990. "Subgroup Analysis." *Biometrics* 46(4):1227–1230.

Butler, Daniel M. and David E. Broockman. 2011. "Do Politicians Racially Discriminate Against Constituents? A Field Experiment on State Legislators." *American Journal of Political Science* 55(3):463–477.

Carvalho, C, N Polson and J Scott. 2010. "The Horseshoe Estimator for Sparse Signals." *Biometrika* 97:465–480.

Chernozhukov, Victor, Denis Chetverikov, Esther Demirer, Mertand Duflo, Christian Hansen, Whitney Newey and James Robins. 2018. "Double/Debiased Machine Learning for Treatment and Structural Parameters." *The Econometrics Journal* .

Esarey, Justin and Jane Lawrence Summer. 2015. "Marginal Effects in Interaction Models: Determining and Controlling the False Positive Rate." Working Paper.

Gelman, Andrew and Eric Loken. 2014. "The Statistical Crisis in Science." *American Scientist* 102(6):460–465.

Green, Donald P. and Holger L. Kern. 2012. "Modeling heterogeneous treatment effects in survey experiments with Bayesian Additive Regression Trees." *Public Opinion Quarterly* 76:491–511.

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):1–22.

Hainmueller, Jens, Daniel J Hopkins and Teppei Yamamoto. 2014. "Causal inference in conjoint analysis: Understanding multidimensional choices via stated preference experiments." *Political Analysis* 22(1):1–30.

Hastie, Trevor, Robert Tibshirani and Jerome Friedman. 2013. *The Elements of Statistical Learning.* 10 ed. New York: Springer-Verlag.

Hill, Daniel and Zachary Jones. 2014. "An Empirical Evaluation of Explanations for State Repression." *American Political Science Review* 108(3):661–687.

Hill, Jennifer, Christopher Weiss and Fuhua Zhai. 2011. "Challenges With Propensity Score Strategies in a High-Dimensional Setting and a Potential Alternative." *Multivariate Behavioral Research* 46(3):477–513.

Holland, Paul W. 1986. "Statistics and Causal Inference (with Discussion)." *Journal of the American Statistical Association* 81:945–960.

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

Imbens, Guido W. and Donald B. Rubin. 2015. *Causal inference for statistics, social, and biometical sciences.* Cambridge University Press.

Lagakos, Stephen W. 2006. "The Challenge of Subgroup Analyses: Reporting without Distorting." *New England Journal of Medicine* 354:1667–1669.

LeBlanc, M. and C Kooperberg. 2010. "Boosting predictions of treatment success." *Proceedings of the National Academy of Sciences* 107:13559–13560.

Montgomery, Jacob M. and Santiago Olivella. 2018. "Tree-based models for political science data." *American Journal of Political Science* .

Murphy, Susan A. 2003. "Optimal Dynamic Treatment Regimes (with discussions)." *Journal of the Royal Statistical Society, Series B (Statistical Methodology)* 65(2):331–366.

Neyman, Jerzy. 1979. "C(α) Tests and Their Use,." *Sankhya: The Indian Journal of Statistics* 41:1–21.

Park, Trevor and George Casella. 2008. "The bayesian lasso." *Journal of the American Statistical Association* 103(482):681–686.

Ratkovic, Marc and Dustin Tingley. 2017. "Sparse Estimation and Uncertainty with Application to Subgroup Analysis." *Political Analysis* 1(25):1–40.

Robinson, Peter. 1988. "Root-N Consistent Semiparametric Regression." *Econometrica* 56(4):931–954.

Rothwell, Peter M. 2005. "Subgroup analysis in randomized controlled trials: importance, indications, and interpretation." *The Lancet* 365(9454):176–186.

Samii, Cyrus. 2016. "Causal Empricism in Quantitative Research." *Journal of Politics* 78(3):941–955.

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):1359–66.

van der Vaart, Aad. 1998. *Asymptotic Statistics.* Vol. 3 of *Cambridge Series in Statistical and Probabilistic Mathematics* Cambridge University Press.

Wager, Stefan and Susan Athey. 2017. "Estimation and Inference of Heterogeneous Treatment Effects using Random Forests." *Journal of the American Statistical Association* .

Wasserstein, Ronald L. and Nicole A. Lazar. 2016. "The ASA's Statement on p-Values: Context, Process, and Purpose." *The American Statistician* 70(2):129–133.

White, Ariel, Anton Strezhnev, Christopher Lucas and Dominika Kruszewska. 2018. "Investigator Characteristics and Respondent Behavior in Online Surveys." *Journal of Experimental Political Science* 5(1):56–67.