Reducing Model Misspecification and Bias in the Estimation of Interactions*

Matthew Blackwell† and Michael Olson‡

October 30, 2019
Word Count: 9,118

Abstract

Studying variation in treatment effects across subsets of the population is an important way that scholars in the social sciences evaluate theoretical arguments. A common strategy to assess such treatment effect heterogeneity is to include a multiplicative interaction term between the treatment and a hypothesized effect modifier in a regression model. In this paper, we show that this approach results in biased inferences due to unmodeled interactions between the effect modifier and other covariates. Researchers can avoid bias by including these additional interactions, but this can lead to unstable estimates due to overfitting. We propose an alternative strategy that uses machine learning techniques to greatly reduce the bias in estimating interactions while guarding against large increases in uncertainty. Simulation evidence shows that our approach outperforms traditional methods for estimating interactions. Finally, we show in two empirical examples that the choice of method leads to dramatically different conclusions about effect heterogeneity.

*Thanks to Stephen Chaudoin, Kosuke Imai, Josh Kertzer, Gary King, Horacio Larreguy, Christoph Mikulaschek, Pia Raffler, Maya Sen, Daniel Smith, Dustin Tingley, Yuhua Wang, Soichiro Yamauchi, and Xiang Zhou for helpful comments and discussions. Open source software to implement the method of this paper are included in the `inters` R package. All errors remain our own.

†Department of Government, Harvard University. email: mblackwell@gov.harvard.edu, web: http://www.mattblackwell.org.

‡Department of Government, Harvard University. email: michaelolson@g.harvard.edu, web: http://www.michaelpatrickolson.com.

1 Introduction

The social and political worlds are full of heterogeneity. Exploring such heterogeneity in treatment effects has become an important and widely used aproach in applied social science research. Indeed, examining varying treatment effects allows scholars to evaluate competing theories about social science phenomena and to better understand mechanisms behind some causal effect. For example, knowing that a particular framing of a political news story affects Democrats differently than Republicans clarifies how framing, information, and partisanship work together to shape the political world. Seeing an effect of remittances on political protest in non-democracies but not in democracies rules out potential mechanisms that would be common to both types of countries. Reliable estimates of effect heterogeneity may also help decisionmakers target their efforts to achieve the most positive impact.

The standard approach to testing these hypotheses is to add a single multiplicative interaction between the main variable of interest and the hypothesized moderator to a "baseline" regression model. A large and growing literature in political methodology has helped clarify these estimands with a particular focus on interpretation, visualization, and sensitivity to hidden assumptions (Braumoeller, 2004; Brambor, Clark, and Golder, 2006; Franzese and Kam, 2009; Berry, DeMeritt, and Esarey, 2010; Kam and Trussler, 2017; Esarey and Sumner, 2018; Bansak, 2018; Hainmueller, Mummolo, and Xu, 2019; Beiser-McGrath and Beiser-McGrath, 2019). Together, these studies have dramatically improved applied researchers' use and presentation of models with interaction effects. Most of these papers, however, focus on situations where, aside from the interaction itself, the regression model is correctly specified.

In this article, we build on this literature and focus on a key potential problem in estimating interaction effects: how the misspecification of "base effects" of the moderator can lead to dramatically biased estimates of the treatment-moderator interaction. In particular, we show how adding a single treatment-moderator interaction to a regression model implicitly assumes no additional interactions between the moderator and other covariates in the model. If the relationship between the covariates and the outcome also depends on the moderator, a naive application of the single-interaction model

can lead to what we call *omitted interaction bias*, a form of model misspecification that we show can be severe. We argue that this type of moderator-covariate interaction is likely to hold in observational data but often goes unnoticed by applied researchers. This source of bias has been noted in a handful of papers in statistics and political methodology (Vansteelandt et al., 2008; Beiser-McGrath and Beiser-McGrath, 2019) but is only rarely discussed or addressed in applied political science research. We build on the methodological work on this type of bias to show how these types of modeling choices intersect with and are distinct from the concept of causal effect identification, clarifying what causal quantity of interest these types of regressions are targeting.

If single interaction terms can create such bias, what alternative do applied researchers have? When the moderator is binary, we show that one simple approach that avoids this omitted interaction bias is to run the baseline model within each level of the moderator and compare the coefficients on treatment across these models. An analogous approach that extends to the non-discrete moderator case is to simply interact the moderator with treatment *and* all covariates in what we call a "fully moderated model." For applied researchers interested in checking the robustness of their single-interaction model point estimates to more flexible specifications, this fully interacted approach may be sufficient. Unfortunately, however, this fully moderated approach can lead to overfitting of the regression model when there are many covariates, possibly leading to unstable estimates and large standard errors.

To avoid these problems, we also develop a data-driven approach that uses machine learning techniques to select which interactions can best combat against bias for the treatment-moderator interaction. Intuitively, the goal of this approach is to use machine learning to choose the "correct" covariate-moderator interactions, thus guarding against both the bias endemic to the single-interaction model and the inefficiency of the fully moderated model. To do so, we use an adapted version of the lasso, or L_1-regularization, a popular technique for prediction that produces *sparse* models, or models that have many estimated coefficients set to zero. While others have suggested machine learning for this type of model misspecification (Beiser-McGrath and Beiser-McGrath, 2019), application of the standard lasso has two flaws for the present setting. First, the retained coefficients

from the lasso are known to be biased, a feature known as *regularization bias*. Second, the lasso applied to just the outcome model may fail to select variables that are important for the independent variable of interest (here, the treatment-moderator interaction), which causes bias. To address both of these issues, we adapt the post-double-selection approach of Belloni, Chernozhukov, and Hansen (2014) to this problem. This approach solves the first problem by only using the lasso for model selection, not estimation; it solves the second problem by using the lasso on both the outcome *and* the treatment-moderator interaction and taking the union of variables selected by those models as the conditioning set. This approach allows us to, essentially, trade off some statistical efficiency (relative to the single-interaction model) to guard against large biases due to model misspecification.

This paper joins studies such as Brambor, Clark, and Golder (2006), Franzese and Kam (2009), Hainmueller, Mummolo, and Xu (2019), and Beiser-McGrath and Beiser-McGrath (2019) in offering applied researchers easy-to-implement solutions to potentially serious problems encountered when estimating and interpreting interactive regression models. Our goal, in this spirit, is to offer intuition and estimators that applied researchers can readily use to execute their already-planned analyses of interactive effects. Our paper is most closely related to Beiser-McGrath and Beiser-McGrath (2019), a recent paper that raises some of the key points about bias we do and uses simulations to assess the performance of various machine learning methods in this setting. We differ from their approach in selecting a machine learning method, post-double selection, that sidesteps many of the problems with machine learning listed above and provides straightforward measures of uncertainty such as standard errors.

We focus on methods for estimating a particular interaction of theoretical interest, and so our recommendations are suited for "confirmatory" analyses, rather than "exploratory" analyses that seek to find interesting heterogeneous effects with limited theoretical guidance. A confirmatory approach usually focuses on a low-dimensional quantity of interest—in our case a main effect and an interaction—and views the high-dimensional covariate space as a nuisance. In developing our intuition and solutions to omitted interaction bias, however, we draw on a broad literature using machine learning to characterize the heterogeneity of treatment effects in terms of some subset of the high-

dimensional covariates (Imai and Ratkovic, 2013; Ratkovic and Tingley, 2017; Künzel et al., 2019). Such a characterization can help target future treatments to the particular units that are most responsive or help explore the possible causal mechanisms at work for a particular treatment effect, but they may or may not be informative about a particular quantity of interest. Thus, while our use of machine learning overlaps with these studies, we differ by focusing on the confirmation of a hypothesized interaction.

Our article proceeds as follows. First, we describe the basic setting and formally demonstrate how model misspecification for interaction can occur. We do so in the common and straightforward case of linear regression, and also in a nonparametric setting that allows us to clearly define causal quantities of interest. We then describe the fully moderated model and the post-double-selection approach, including our extension of this estimator to fixed effects models and settings with clustered errors. We demonstrate the relative strengths of different estimation approaches using a simulation study, and show the potential importance of the issue using two empirical illustrations. We conclude with thoughts about best practices with interaction terms.

2 The Problem

2.1 Multiplicative Interactions in Linear Models

Suppose we have a random sample from a population of interest labeled $i = 1, \ldots, N$. For each unit in the sample we measure the causal variable of interest, or treatment, D_i, an outcome Y_i, a potential moderator V_i, and a $K \times 1$ vector of additional controls, X_i. In particular, we are interested in how the effect of D_i on Y_i varies across levels of V_i, controlling for the additional covariates, X_i. We consider the following "base" regression model that a researcher might use to assess the effect of treatment:

$$Y_i = \alpha_0 + \alpha_1 D_i + \alpha_2 V_i + X_i' \alpha_3 + \varepsilon_{i1} \tag{1}$$

A common way to assess treatment effect heterogeneity is to augment this model with a single multiplicative interaction term between treatment and the moderator, which we call the *single-interaction*

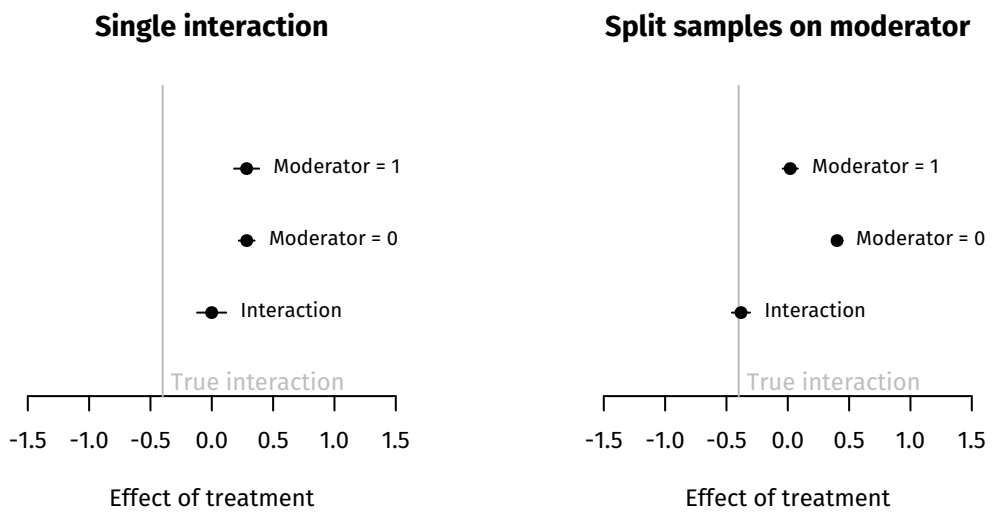

Figure 1: An simulated example of model misspecification in interaction models.

model:

$$Y_i = \beta_0 + \beta_1 D_i + \beta_2 V_i + X_i' \beta_3 + \beta_4 D_i V_i + \varepsilon_{i2}, \tag{2}$$

where β_4 is the quantity of interest.

An alternative estimation strategy that may, at first glance, appear equivalent to (2) is to estimate the base model (1) within levels of V_i (obviously omitting the $\alpha_2 V_i$ term). From standard results on the linear regression, these two approaches will be equivalent when there are no additional covariates, X_i, in these models. When those covariates are present, however, they can differ substantially. Figure 1 shows a simulated example of this in action, with a single X_i, and binary D_i and V_i (the full simulation code is available in the replication archive). Here, we see that when running the single-interaction model (2), it appears as if there is no effect heterogeneity across levels of V_i, but when we split the sample on V_i, there is a large and meaningful difference in effects, one that aligns with the true value of the interaction.

Why does the split-sample approach capture the true interaction effect in this case when the single-interaction model cannot? It is helpful to note that the split sample approach is equivalent to running a *fully moderated model*, where V_i is interacted with *all* of the variables:

$$Y_i = \delta_0 + \delta_1 D_i + \delta_2 V_i + X_i' \delta_3 + \delta_4 D_i V_i + V_i X_i' \delta_5 + \varepsilon_{i3} \tag{3}$$

If this model represents the true data-generating process, then using ordinary least squares to estimate the single-interaction model will result in a biased estimator for the interaction of interest, $\widehat{\beta}_4$. Under the standard omitted variable bias formula, we have

$$\widehat{\beta}_4 \xrightarrow{p} \delta_4 + \gamma_v' \delta_5,$$

where γ_v is the population regression coefficients of the $V_i X_i$ interactions on $D_i V_i$, controlling for the other variables in the single-interaction model. Thus, the single-interaction model can produce misleading estimates when (a) the treatment-moderator interaction is predictive of the omitted interactions, and (b) the omitted interactions are important for predicting the outcome. Thus, an estimated interaction from a single-interaction model could be due to the moderator as hypothesized or due to some unmodeled heterogeneity in the interactive effects. We refer to this possible bias, $\gamma_v' \delta_5$, as *omitted interaction bias*. Note that the inclusion of treatment-covariate interactions ($D_i X_i$) does not fully address this issue, because these do not account for interactions between the moderator and the covariates.

Intuitively, this type of omitted interaction bias occurs because the covariates have different relationships with the outcome across levels of the moderator. In the split-sample or fully moderated approaches, this variation in the conditional relationship between X_i and Y_i is allowed, whereas in the single interaction model, it is assumed away. Thus, even if a scholar is convinced that they have chosen the correct model for the baseline regression, hypothesized moderators pose a new challenge. In particular, there is an awkward tension in assuming that a potential moderator is important enough to test for interactive effects with treatment, but simultaneously assume it does not also moderate other covariates. There are a few settings where we might expect this omitted interaction bias to be zero. In particular, there will be no such bias when treatment D_i, the moderator V_i, and covariates X_i are all randomized, as would be the case in a factorial or conjoint experiment. In those cases, $\gamma_v = 0$ and so there will be no omitted interaction bias. Thus, our discussion here most closely applies to situations where X_i represents a set of observational controls where independence will almost certainly be violated.

While there has been much written about interactions in political science, little of that literature has focused on these omitted interactions. Early efforts advocated for inclusion of base terms and the use of plots to assess marginal effects (Braumoeller, 2004; Brambor, Clark, and Golder, 2006) and correct interpretation of base terms (Franzese and Kam, 2009); more recent work studies interactions in non-linear models (Berry, DeMeritt, and Esarey, 2010), sensitivity to linearity assumptions in interactive regression models (Hainmueller, Mummolo, and Xu, 2019), and causal identification in contexts without a randomized moderator (Bansak, 2018). This paper builds on these studies, as most have taken the single-interaction model as given and, at least implicitly, correctly specified.[1] Beiser-McGrath and Beiser-McGrath (2019) is one of the few studies to address this issue directly and focuses on interactions and nonlinear terms in parametric regressions, though our recommendations differ in terms of the machine learning approach used to avoid this bias (which we discuss further below). In addition, we extend this literature by connecting these issues of bias to the nonparametric identification of causal effects in the next section.

2.2 Nonparametric Analysis and Interactions as Modeling Assumptions

While a linear regression context is perhaps the most intuitive—and immediately useful—way to understand the omitted interaction bias issue, most scholars use linear regression not as an end in itself but rather as a tool to estimate causal inferences about social and political phenomena. Thus, it is valuable to define our causal quantities of interest and assumptions in a nonparametric setting.

We take the view of a researcher interested in the causal effect of D_i and how that causal effect varies by the effect modifier V_i. Let $Y_i(d)$ be the potential outcome for unit i when treatment is at level d, so that an average treatment effect may be defined as

$$\tau(d, d^*) = \mathbb{E}[Y_i(d) - Y_i(d^*)].$$

We can connect the potential outcomes to the observed outcomes with a consistency assumption that $Y_i = Y_i(d)$ when $D_i = d$. With a binary moderator, we can define the interaction between treatment

[1] We note that the kernel estimator described by Hainmueller, Mummolo, and Xu (2019) can also allow for interactions between the moderator and other covariates, though those authors do not focus on that modeling decision.

and the moderator as:

$$\delta(d, d^*) = \mathbb{E}[Y_i(d) - Y_i(d^*) \mid V_i = 1] - \mathbb{E}[Y_i(d) - Y_i(d^*) \mid V_i = 0]. \tag{4}$$

Note that we are *not* explicitly considering causal interactions (VanderWeele, 2015; Bansak, 2018), wherein the interaction effect are defined in terms of joint potential outcomes, $Y_i(d, v)$, and can itself be interpreted causally. To use these joint counterfactuals, researchers would need to identify both the causal effect of V_i *and* D_i, which is unrealistic in most settings. Thus, we focus on descriptive heterogeneity in an estimated causal effect as measured by (4).

When attempting to estimate these types of causal effects, it is helpful to classify assumptions into two types: identification assumptions and modeling assumptions. Identification assumptions are those that allow us to connect causal (that is, counterfactual) quantities of interest to statistical parameters of an observable population distribution. For instance, a common assumption invoked in observational studies to estimate a causal effect in the above base regression model would be "no unmeasured confounding," or $Y_i(d) \perp\!\!\!\perp D_i \mid V_i, X_i$, where $A \perp\!\!\!\perp B \mid C$ means that A is independent of B conditional on C. Under this identification assumption, we can connect the conditional expectation of the potential outcomes to conditional expectation of the observed outcome:

$$\mathbb{E}[Y_i(d) \mid V_i, X_i] = \mathbb{E}[Y_i \mid D_i = d, V_i, X_i].$$

Thus, the interaction between D_i and V_i is nonparametrically identified as

$$\begin{aligned}\delta(d, d^*) = &\int_{x \in \mathscr{X}} (\mathbb{E}[Y_i \mid D_i = 1, V_i = 1, X_i = x] - \mathbb{E}[Y_i \mid D_i = 0, V_i = 1, X_i = x])\, dF_{X|V}(x|V_i = 1) \\ &- \int_{x \in \mathscr{X}} (\mathbb{E}[Y_i \mid D_i = 1, V_i = 0, X_i = x] - \mathbb{E}[Y_i \mid D_i = 0, V_i = 0, X_i = x])\, dF_{X|V}(x|V_i = 0),\end{aligned} \tag{5}$$

where $F_{X|V}(x|v)$ is the distribution function of X_i given V_i. This result is nonparametric in the sense that it places no restrictions on the joint distribution of the observed data. In particular, the interaction is identified from the data before we make any assumptions about what interaction terms "belong" in the regression models. Omitted variable bias usually refers to the case when no unmeasured confounding (the key identification assumption) is incorrect, but there is an additional variable, Z_i, that if added to X_i, would ensure that the assumption would hold.

Modeling assumptions, on the other hand, place restrictions on the observable population distribution. For example, linearity of the observable conditional expectation function (CEF), $\mathbb{E}[Y_i|D_i, V_i, X_i]$, is a modeling assumption because it places restrictions on the conditional relationship between X_i and Y_i. Other modeling assumptions include homoskedastic error variances and conditionally normal errors. These assumptions are often made for statistical reasons such as efficiency since many estimators are more efficient under stronger modeling assumptions. The various assumptions about interactions in the above models are modeling assumptions and imply simplified expressions for the quantity $\delta(d, d^*)$. For instance, under the base regression model, we have $\delta(d, d^*) = 0$, whereas in the single-interaction model we have $\delta(d, d^*) = \beta_4 \times (d - d^*)$, and in the fully moderated model we have $\delta(d, d^*) = \delta_4 \times (d - d^*)$. These formulations of the interactions become more complicated (and will depend on X_i) when the models for the CEF contain interactions between treatment and covariates, though all the points about moderator-covariate interactions remain. When these modeling assumptions are incorrect, we call this *model misspecification* and note that it can cause bias, just as omitted variables do in terms of identification. Finally, we note that in many cases, different modeling assumptions can be nested in the sense that a more flexible model can contain a more constrained model as a subset. In our example, the fully moderated model reduces to the single-interaction model when $\delta_5 = 0$, and so the former can represent the latter but not the reverse.

Why distinguish between these types of assumptions? Both identification assumptions and modeling assumption can be violated in practice, and both kinds of violations can lead to bias or inconsistency in the estimation of the quantity of interest. Identification assumption, though, cannot be verified or falsified directly by the data, whereas modeling assumptions can be tested by comparing a given model to a more flexible one. This means that while violations of the identification assumption preclude any causal estimation, model misspecification can in principle be eliminated by always using more flexible models—that is, those that encode fewer modeling assumptions. We can make the single-interaction model more flexible by simply including moderator-covariate interactions into our model, and, thus in terms of bias, we should prefer the fully moderated model. It is better able to produce an accurate approximation to the underlying conditional expectation function of interest,

$\mathbb{E}[Y_i|D_i, V_i, X_i]$. Of course, the reduction of bias comes at the cost of increased uncertainty, so below we demonstrate how machine learning techniques can be used to choose which interactions in the fully moderated model are important for estimating the treatment-moderator interaction and which can be abandoned for efficiency's sake.

Finally, we note that the choice of modeling assumptions is sometimes confused with the choice of quantity of interest. For example, researchers often use the above base regression that omits a interaction between D_i and V_i in part because they are targeting the *average* or *overall* effect of treatment. They then turn to alternative modeling assumptions—those encoded in the single-interaction model—when their quantity of interest changes to the effect heterogeneity of D_i across V_i. We note that this practice, while commonplace, is not required since researchers can use interaction models such as the single-interaction and fully moderated models to recover average treatment effects even though they are not (necessarily) encoded in a single parameter of the model. Thus, many of the same modeling decisions we discuss here could also be used when targeting the average treatment effect. Indeed, previous work has emphasized that running separate regression models for treatment and control groups (and implicitly including treatment-covariate interactions) is a good way to estimate the overall effect (Imbens, 2004). The specific choice of X_iV_i interactions, though, is often more consequential for estimation of the D_iV_i interaction (rather than the main effect of D_i) because of the inclusion of V_i in both multiplicative terms.

3 Flexible Estimation Methods for Interactions

How can scholars avoid the misspecification of the single-interaction model? We highlight two possibilities that offer a combination of easy implementation and interpretation, while addressing the omitted interactions problem. While much of the discussion in this paper revolves around the moderator-covariate interactions, both of the approaches outlined below can also incorporate treatment-covariate interactions or even covariate nonlinearities in a straightforward manner.

3.1 Fully Moderated Models

As discussed above, the most straightforward strategy for avoiding the misspecification of the single-interaction model is to simply estimate the fully moderated model, (3). This is equivalent to split-sample estimation when the moderator is binary, but allows for other types of moderators as well. For full flexibility, the moderator must be interacted not only with observable covariates, but also with controls for unobserved unit or time fixed effects, if they are included in the model. The estimation and interpretation of the marginal effects of the treatment and the interaction remain similar to the single-interaction model (2).

3.2 Post-double Selection

One concern with a fully moderated model is the dramatic proliferation of parameters that it generates. Adding an interaction between the moderator and all covariates will nearly double the number of parameters to be estimated in the model, which is problematic in models with large numbers of covariates or fixed effects. This is in addition to any treatment-covariate interactions or nonlinearities a researcher wants to include. As in any scenario with a large number of parameters relative to sample size, this can generate both a loss of statistical power as well as overfitting.

As a solution to these concerns, we propose using a standard approach to guarding against overfitting: regularization through penalized parameter estimation. Specifically, we perform variable selection and estimation of model parameters through use of a lasso (Tibshirani, 1996). The lasso is a penalized regression procedure that induces sparsity so that many of the coefficients are estimated to be precisely zero. This procedure is very attractive for model selection, but it unfortunately has a few disadvantages that make it ill-suited for our task at hand. First, the standard lasso guards against overfitting by shrinking coefficients toward zero, which could lead to significant bias in effect estimates for retained coefficients, even asymptotically (Knight and Fu, 2000). This so-called *regularization bias* may be beneficial when attempting to optimize prediction accuracy, but here we focus on the estimation of a particular effect or a particular interaction. Second, in finite samples, the naive lasso can make costly model selection "mistakes" by setting small coefficients to zero when

they are strongly correlated with, say, the treatment. This can lead to large omitted variable biases for the treatment since that bias depends on both the covariate-outcome relationship *and* the covariate-treatment relationship (Belloni, Chernozhukov, and Hansen, 2014). This same issue applies to the treatment-moderator interaction. Other forms of variable selection and regularization, such as the Bayesian lasso (Park and Casella, 2008) or elastic net methods (Zou and Hastie, 2005), inherit some of these issues.

To avoid the biases of the standard lasso and to perform inference on the key quantities of interest, we apply the post-double selection (PDS) procedure of Belloni, Chernozhukov, and Hansen (2014). Unlike the standard lasso, this estimator takes the estimation of treatment effects or some other low-dimensional parameter as its explicit goal, making it ideally suited to our application. This procedure uses the lasso with data-dependent and covariate-specific penalties that select a subset of variables that can well-approximate the conditional expectation function of interest. It does so by applying the lasso to not only the outcome, but also the main *independent* variables of interest (here, D_i and $D_i V_i$) to find variables that predict any of these variables well. Finally, the union of the selected variables is passed to a standard least-squares regression with possibly heteroskedastic- or cluster-robust standard errors. By using the union of variables selected to predict both the outcome and the independent variables of interest well (the "double selection" in PDS), this procedure minimizes the potential for omitted variable bias due to incorrect model selection by the lasso. And by using standard OLS for the final estimation after theses lasso steps (the "post" in PDS), we avoid the regularization bias of the standard lasso.

To apply the PDS approach to the current setting, we take the main effect D_i and the interaction $D_i V_i$ as the main variables of interest and let $Z_i' = [V_i \; X_i' \; V_i X_i']$ be the vector of remaining variables from the fully moderated model (where we assume they have been mean centered). We then run lasso regressions with each of $\{Y_i, D_i, D_i V_i\}$ as dependent variables and Z_i as the independent variables in each model, using the data-driven penalty loadings suitably adjusted for the clustering in our applications (Belloni et al., 2016).

$$\widehat{\gamma}_y = \arg\min_{\gamma_y} \sum_{i=1}^{N} (Y_i - Z_i'\gamma_y)^2 + \sum_{j=1}^{k} \lambda_{yj}|\gamma_{yj}| \tag{6}$$

$$\widehat{\gamma}_d = \arg\min_{\gamma_d} \sum_{i=1}^{N} (D_i - Z_i'\gamma_d)^2 + \sum_{j=1}^{k} \lambda_{dj}|\gamma_{dj}| \tag{7}$$

$$\widehat{\gamma}_{dv} = \arg\min_{\gamma_{dv}} \sum_{i=1}^{N} (D_i V_i - Z_i'\gamma_{dv})^2 + \sum_{j=1}^{k} \lambda_{dvj}|\gamma_{dvj}| \tag{8}$$

Let Z_i^* be a vector of the subset of Z_i that has either $\widehat{\gamma}_y$, $\widehat{\gamma}_d$, or $\widehat{\gamma}_{dv}$ not equal to zero. The final step of post-double selection is to regress Y_i on $\{D_i, D_i V_i, Z_i^*\}$ using OLS.

Belloni, Chernozhukov, and Hansen (2014) showed that, under regularity conditions, this procedure will give consistent estimates of the coefficients of interest and the standard robust or cluster-robust sandwich estimators for the standard errors will be asymptotically correct. The key regularity condition of this approach is *approximate sparsity*, which states that the conditional expectation functions of each of the outcomes given Z_i can be well-approximated by a sparse subset of Z_i and that the size of this sparse subset grows slowly relative to the sample size. This is a considerably weaker condition than the usual exact sparsity requirement of the lasso, where many of the covariates must have exact zero coefficients. This assumption also fits well with the context of moderator-covariate interactions, which we might be willing to believe are mostly small, though not so confident as to say they are exactly zero.

The penalty loadings in the lasso selection models vary by both the outcome in the lasso and the covariate. In order to achieve consistency and asymptotic normality, these loadings must be chosen carefully. Belloni, Chernozhukov, and Hansen (2014) show that the ideal penalty loadings are a function of the interaction between the covariates and the error for that outcome. For instance, for the outcome we have $\lambda_{yj} = \lambda_{y0}\sqrt{(1/N)\sum_{i=1}^{N} Z_{ij}^2 \varepsilon_{yi}^2}$. Intuitively, this regularize variables more if their "noise" correlates with the error. These infeasible loadings can be estimated using a first-step lasso to provide estimates of the error, $\widehat{\varepsilon}_{yi}$, as with robust variance estimators. Belloni, Chernozhukov, and Hansen (2014) show that this procedure (along with a carefully chosen value of the λ_{y0}) will

allow for consistency and asymptotic normality even when the errors are non-normal and possible heteroskedastic.

It is possible to override the lasso selections and force the inclusion of some variables in the final model. In the empirical examples below, we force V_i and X_i to be included in the final model selection, regardless of how the lasso estimates their coefficients. This helps isolate the change in the estimated interactions due to interaction modeling alone and ensures that the original model for the marginal effect of D_i is nested in the model for effect heterogeneity. A second benefit of this modeling choice is that it avoids a situation where the lasso estimates base terms of, say, X_{ij} is zero, but selects the interaction $V_i X_{ij}$ to be included in the model. An alternative, and much more general approach, is the hierarchical lasso, which performs variable selection on *all* possible bivariate interactions while remaining sensitive to the hierarchical character of interaction estimation (Bien, Taylor, and Tibshirani, 2013).

Finally, Beiser-McGrath and Beiser-McGrath (2019) also suggest regularization methods to address concerns about overfitting and efficiency loss, and investigate several possibilities including the adaptive lasso, kernel regularized least squares, and Bayesian additive regression trees. Of these, only the adaptive lasso has been shown to overcome the regularization bias described above, though it does not avoid the problem of costly model selection mistakes for potential confounders. One additional advantage of the post-double-selection approach is that it provides straightforward estimates of uncertainty without using resampling methods such as the bootstrap, which may have poor performance with the lasso if used without modifications (Camponovo, 2015).

3.3 Fixed Effects and Clustering with the Lasso

One source of substantial numbers of parameters in many regression models is unit or time fixed effects. For the base regression model, these factors can be incorporated without having to estimate additional parameters by various demeaning operations. For fully interacted model, on the other hand, they must be included as interactions between a binary variable representation of the units or time periods (usually omitting a reference category) and the moderator. But this may add a signif-

	R_{i1}	R_{i2}	R_{i3}
Northeast	-1	-1	-1
Midwest	1	0	0
West	0	1	0
South	0	0	1

Table 1: Deviation coding example

icant number of parameters to the model, and so it may be fruitful to regularize those interactions. Unfortunately, the typical dummy variable representation of fixed effects is poorly suited for regularization. Imagine, for instance, that we had a variable for region of the U.S. in our model, with levels {Northeast, Midwest, West, South}. In a typical regression model, we would include dummy variables for, perhaps, Midwest, West, and South, and the coefficients on these dummy variables would be comparisons of the (conditional) average outcomes in each of these categories against the omitted category, Northeast. Thus, shrinking coefficients toward zero in this case means making each region closer to the Northeast region. If there aren't many regions close to the omitted category, then the lasso will not take advantage of its sparsity.

Instead of this typical reference or dummy coding of categorical variables, we recommend deviation or sum coding. To illustrate how this coding works, we take the same census region variable and represent it with a series of variables, (R_{i1}, R_{i2}, R_{i3}), that are similar to the typical dummy variable representation of the {Midwest, West, South} regions, except that in each variable, any observation from the omitted category, Northeast, is coded as -1. We show how each variable codes each category in Table 1. The benefit of this coding is that the coefficients on each of these variables has the interpretation of the difference in (conditional) means between each region and the grand (conditional) mean of the groups. Thus, shrinkage toward zero in this case implies shrinkage of each group toward the grand mean, a far more meaningful baseline than an arbitrary omitted category. And while this discussion focused on "main effects," the same reasoning applies to the types of interactions we consider in this paper.

Finally, clustering of units is a common concern in applied work, and scholars often rely on

cluster-robust standard errors to ensure proper uncertainty estimates. Clustering also complicates the PDS approach through the choice of the penalty terms. Belloni et al. (2016) show that a small modification to the penalty will ensure the post-double selection will continue to be produce consistent and asymptotically normal in this setting. In particular, suppose that we have observations in clusters so that Y_{ig} is observation i in group g, with N_g observations in each group, G groups, and $N = \sum_{g=1}^{G} N_g$ total individuals. Then, we would set the penalty parameter as $\lambda_{yj} = \lambda_{y0}\phi_{yj}$, where

$$\phi_{yj}^2 = \frac{1}{N}\sum_{g=1}^{G}\left(\sum_{i=1}^{N_g} Z_{igj}\varepsilon_{yig}\right)^2.$$

For a feasible estimate of this penalty, we can run an initial lasso to obtain estimates of $\widehat{\varepsilon}_{yig}$. The penalty terms for the other lasso regressions follow similarly. Again, the penalty parameter depends on a measure of the noise in estimating the γ_{yj}, but in this case that noise allows for arbitrary dependence within the clusters (Belloni et al., 2016). The difference between this case and the above standard PDS is similar to the difference between calculating the cluster-robust variance estimator and the heteroskedasticity-robust variance estimator.

4 Simulation Evidence

The theoretical properties of the post-double-selection estimator are asymptotic in nature which is useful insofar as they provide reasonable approximations to performance in finite samples. In this section, we describe the results of a Monte Carlo analysis of this approach and several alternative approaches to see how they perform in a variety of finite sample settings. We follow a similar approach to Belloni, Chernozhukov, and Hansen (2014) and draw a set of covariates X_i of dimension K, from $\mathcal{N}(0, \Sigma)$, where $\Sigma_{ij} = 0.5^{|j-k|}$ so that the covariates depend on each other. We vary the number of covariates between a low-dimensional setting, $K = 20$, and a relatively high-dimensional setting,

$K = 200$. We then generate the moderator, treatment, and outcome with the following:

$$\mathbb{P}[V_i = 1 \mid X_i] = \text{logit}^{-1}\left(\delta_{v|0} + X_i'\boldsymbol{\delta}_{v|x}\right) \tag{9}$$

$$D_i = \delta_{d|0} + 0.25 \times V_i + X_i'\boldsymbol{\delta}_{d|x} + V_i X_i'\boldsymbol{\delta}_{d|vx} + \varepsilon_{id} \tag{10}$$

$$Y_i = \delta_{y|0} + 0.5 \times D_i + 0.25 \times V_i + X_i'\boldsymbol{\delta}_{y|x} + 1 \times D_i V_i + V_i X_b \boldsymbol{\delta}_{y|vx} + \varepsilon_{iy} \tag{11}$$

The parameters of these models are generated under a quadratic decay, so that the jth entry of $\boldsymbol{\delta}_{v|x}$ is $\delta_{v|x[j]} = 2/j^2$. We define the other coefficient vectors similarly:

$$\delta_{d|x[j]} = 2/j^2 \quad \delta_{d|vx[j]} = c_{d|vx}/j^2$$

$$\delta_{y|x[j]} = 2/j^2 \quad \delta_{y|vx[j]} = c_{y|vx}/j^2$$

We vary $c_{d|vx}$ and $c_{y|vx}$ so that the $V_i X_i$ interactions have partial R^2 values of $\{0, 0.25, 0.5\}$. Each of the errors, $\{\varepsilon_{id}, \varepsilon_{iy}\}$, are independent standard normal. Note that this set is not sparse in any of the equations, but it is approximately sparse in the sense of Belloni, Chernozhukov, and Hansen (2014).

We apply several methods to this data generating process. First, we apply both the single-interaction and fully moderated OLS models. Second, we apply two lasso-based approaches. The first of these is a post-lasso approach that only selects variables based on a lasso applied to the outcome, using cross-validation and the one standard error rule to select the complexity parameter. The second lasso-based procedure we apply is the above post-double-selection estimator using the penalty choices given by Belloni, Chernozhukov, and Hansen (2014).

Figure 2 shows the results of these simulations. We omit the single interaction terms from these plots because the strong bias of that approach obscures the relative performance of the other methods. We present the full results in Appendix Figure SM.6. Of the methods we show here, the fully moderated and post-double-selection methods both have similar low levels of bias, which makes sense given the fact that the data was generated from a fully moderated model. And in terms of overall root mean square error (RMSE), the two methods have similar performance with a low number of covariates ($K = 20$), but post-double selection is more efficient with a lower RMSE with a high number of covariates ($K = 200$). Both the cross-validated single lasso and the single-interaction model perform well when there are no $V_i X_i$ in the outcome model ($R^2_y = 0$, leftmost panels), but both also

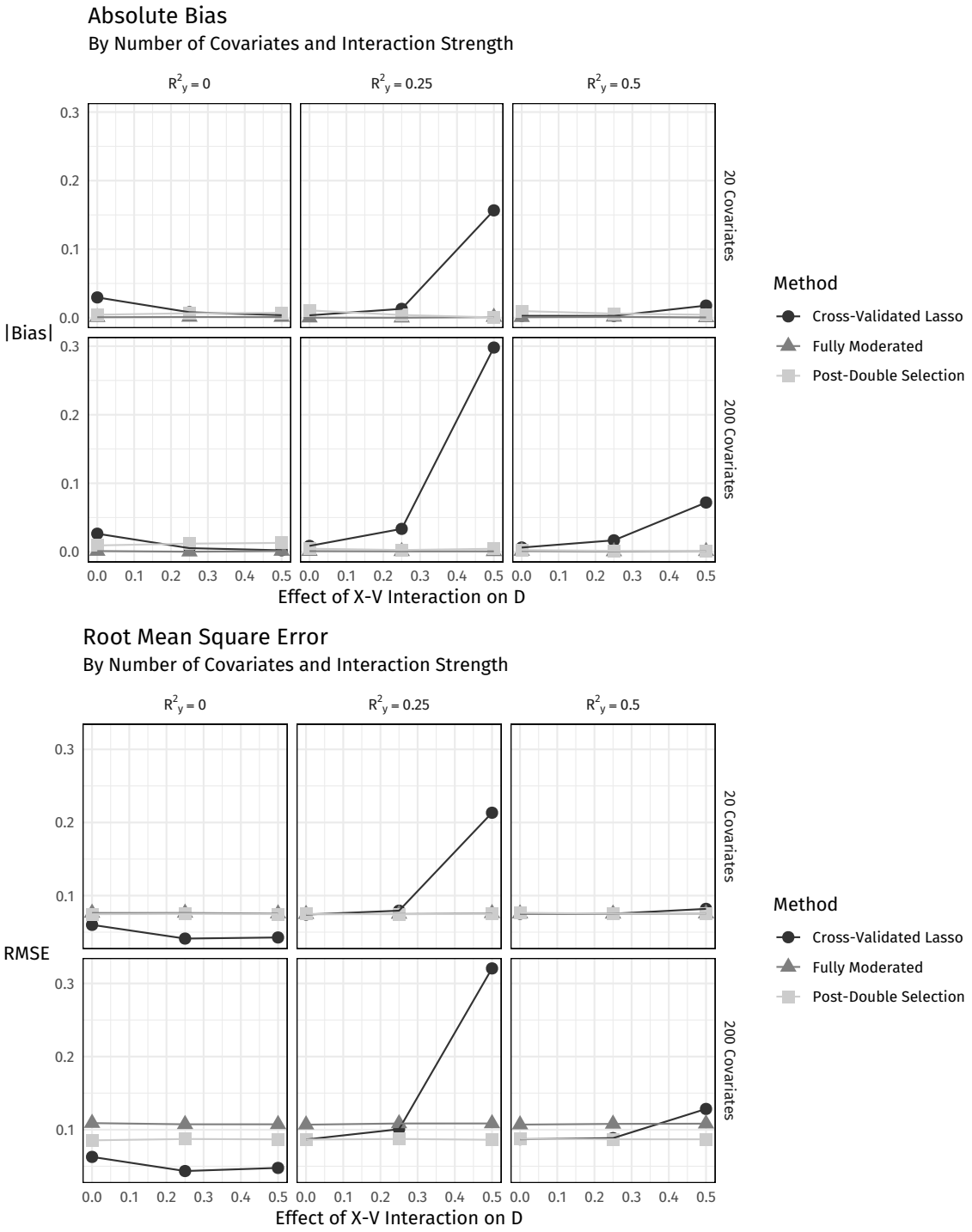

Figure 2: Simulation results

Bias (top) and root mean square error (bottom) of various methods when estimating interactions. Horizontal panels vary the partial R^2 of the $V_i X_i$ interactions on Y_i and vertical panels vary the number of covariates. The x-axis in each panel varies the partial R^2 of the $V_i X_i$ on D_i.

have large degrees of bias when those interactions are present. The performance of the outcome-based cross-validated lasso is worst when $V_i X_i$ are important for D_i, but only moderately important for Y_i. This is when the outcome-based lasso is most likely to make costly selection mistakes that the post-double-selection approach is designed to avoid.

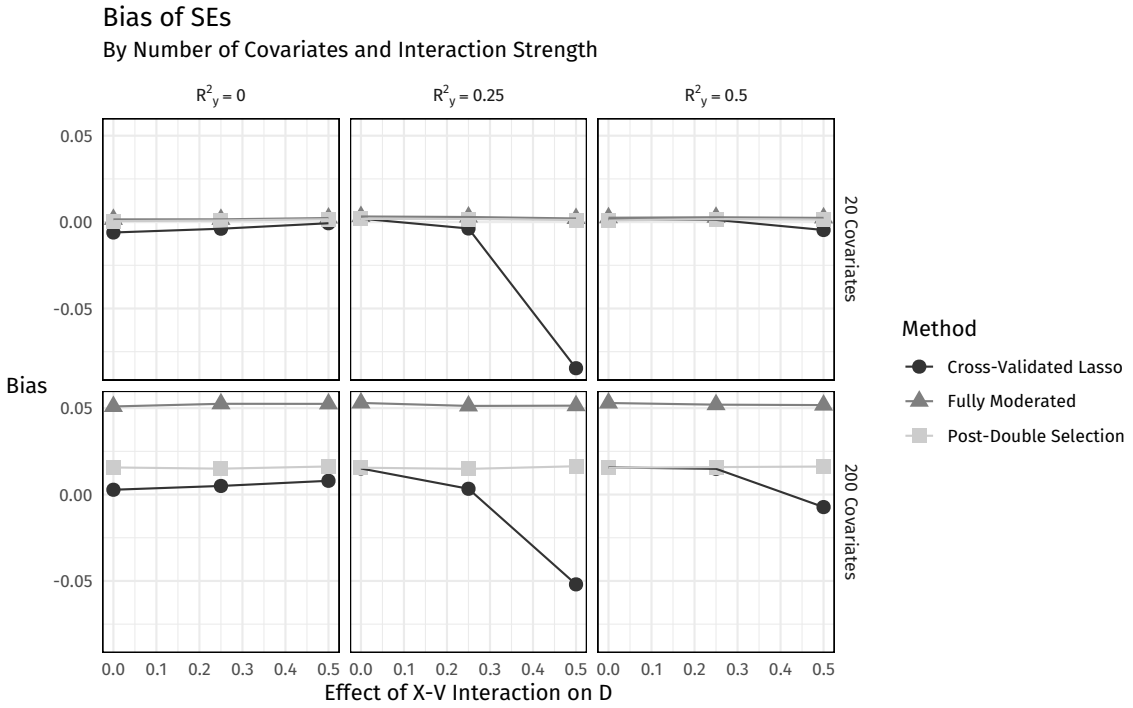

Figure 3: Simulation results for standard error estimation

Bias of the standard error estimates form the various methods when estimating interactions.

Finally, we present the bias of the standard error estimates from these estimators. With a small number of covariates, both the fully moderated and post-double-selection approaches perform extremely well, with the outcome lasso showing downward bias in the standard error estimates across a range of settings. With a high number of covariates, the standard error estimates for the fully moderated model become upwardly biased, whereas the outcome-based lasso show a similar downward bias when the moderator-covariate interactions are important for treatment. Overall, the results on point and variance estimation show that the post-double-selection approach performs well in realistic data situations with a small efficiency loss when the moderator-covariate interactions are completely unrelated to the outcome.

5 Empirical Illustrations

5.1 The Direct Primary and Third-Party Voting

The role of the direct primary in shaping American electoral politics has been of persistent interest to scholars. One argument surrounding this uniquely American institution is that, by creating a clear path to major party nominations by those other than party insiders (Hirano and Snyder, 2007), and by allowing for ideological heterogeneity within parties (Ansolabehere, Hirano, and Snyder, 2007), it reduced the electoral prominence of third parties. This argument is tested directly by Hirano and Snyder (2007) using a two-way fixed effects models to control for state-specific and year-specific unobserved counfounders.

Yet a substantial literature calls into question the similarity of direct primary usage and adoption in the American South and non-South. In the South, the direct primary was part of a suite of reforms designed to disfranchise African Americans (Kousser, 1974; Perman, 2001), and quickly became the *de facto* election in the Democratic Party-dominated southern states. Outside the South, the direct primary was adopted as part of a Progressive effort to increase individuals' voices in government (Merriam, 1908), likely with major parties' implicit consent (Ware, 2002; Reynolds, 2006).

With varying motivations across North and South, it is reasonable to expect—and important to evaluate—whether the effect of direct primary adoption is similar in the two regions.Hirano and Snyder (2007) do so by splitting their sample of states into South and non-South and separately estimating the effect in the two regions—an approach analogous to our fully moderated model and so consistent with our recommendations. We now explore the importance of this modeling decision by comparing fully moderated models such as these with models that allow only the "treatment," direct primary adoption, to vary between regions (single-interaction model).

We focus on U.S. House elections, and take as our outcome variable the share of all U.S. House votes cast in a given state-election for parties or individuals other than Democrats or Republicans.[2] We measure direct primary adoption as an indicator variable for whether the direct primary was in

[2]Data on U.S. House elections is from ICPSR Study 6895, "Party Strength in the United States: 1872-1996."

widespread use in a given state and year.[3] Our moderator, *South*, is an indicator for whether a state is one of the eleven states of the former Confederacy. The simple interaction model can therefore be expressed as the following:

$$(100 - DemShare_{it} - RepShare_{it}) = \beta \left(Primary_{it} \times South_i \right) + \gamma Primary_{it} + \alpha_i + \tau_t + \epsilon_{it} \qquad (12)$$

where i indexes states and t indexes election years. The base term on *South* is absorbed by the state fixed effects α; τ is a year fixed effect. In this straightforward setup, the only interactions added in the fully moderated model are those between year fixed effects and the moderator.

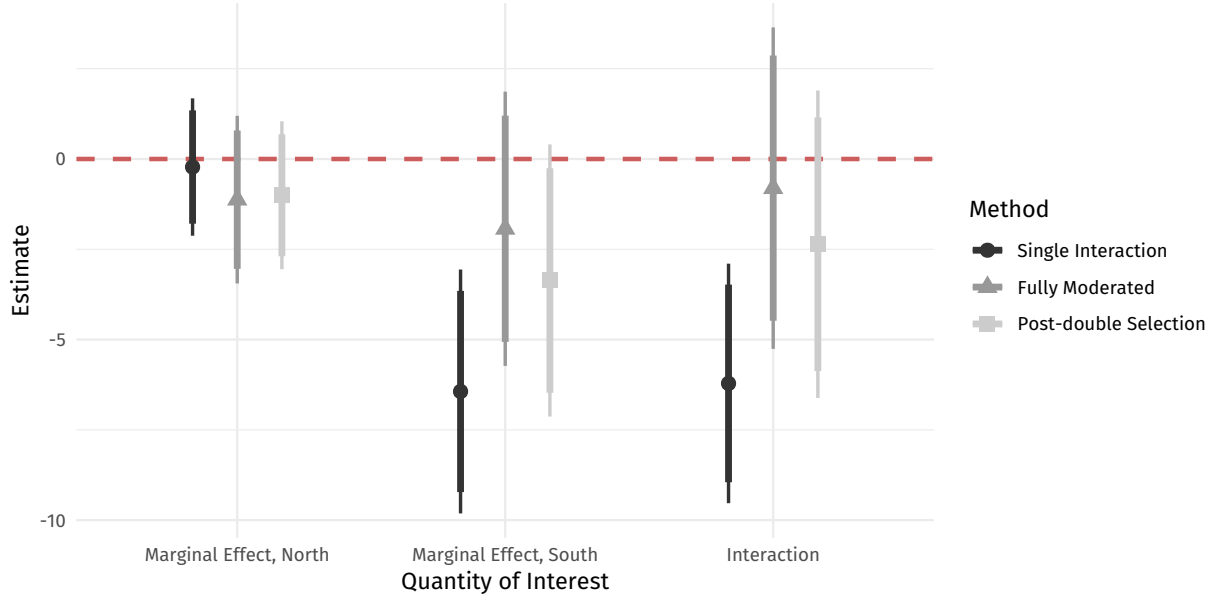

Figure 4: Effect of the Direct Primary in the American North and South

Estimates from the single-interaction, fully moderated, and post-double-selection models described above. 90% and 95% confidence intervals are based on state-clustered standard errors.

Figure 4 displays estimates from a single-interaction model given by Equation (12), a fully moderated model that adds interactions between the year fixed effects and the *South* indicator, and post-double-selection approach that uses a lasso to select over the interactions between the year fixed effects and the *South* indicator interactions. The single-interaction and fully moderated models suggest starkly different conclusions. While these models agree that there is a small, statistically insignificant

[3] We draw this information from Hirano and Snyder (2019), Table 2.A.

effect of direct primary adoption in northern states, the single-interaction model suggests a quite large effect of about six percentage points in southern states; the fully moderated model, on the other hand, suggests only a small, statistically insignificant negative effect. This pattern is repeated in the interaction term coefficients: the fully moderated model suggests little difference between North and South, while the single-interaction model suggests a substantially larger negative effect in the South than the North. The conclusions of the post-double-selection approach are in-between those of the single-interaction and fully moderated models, but are more consistent with the fully moderated model insofar as there is no evidence of a statistically significant difference in the effect of the primary between the two regions. Together, these latter two models call into question the validity of the single-interaction approach and suggest that unmodeled differences in year-specific third-party voting between the North and the South are driving its seemingly significant interaction.

5.2 Regime Type and Remittances

The role of remittances in shaping political activity is an active area of research, with some literature suggesting that remittances can buttress authoritarian governments, and others suggesting that remittances can spur political change in democratizing or non-democratic countries. Entering into this debate, Escribà-Folch, Meseguer, and Wright (2018) explore the relationship between remittances and political protest, a first step on the path of democratization. They argue that remittances ought to be associated with greater levels of protest, but only in non-democracies, and find evidence consistent with this claim.

To do so, the authors use novel (continuous) measures of remittances and protest and an array of control variables in a linear regression model with county and time fixed effects.[4] To test the heterogeneous effects of remittances across regime type, *Remit* is interacted with a binary indicator for regime type, *Autocracy*. This yields the following specification:

$$Protest_{it} = \beta \left(Remit_{it} \times Autocracy_{it} \right) + \gamma Remit_{it} + \phi Autocracy_{it} + \psi' \mathbf{X}_{it} + \alpha_i + \tau_t + \epsilon_{it}, \quad (13)$$

[4]The authors also test their results using an instrumental variables design; we restrict our focus to their main OLS specification.

where $\mathbf{X}_{it}$ is a vector of time-varying controls. In keeping with the above discussion, however, we argue that this model makes important assumptions that can be easily relaxed. Specifically, we note that this model assumes that all covariates—including, importantly, the fixed effects—other than the main treatment of interest have the same (linear) effect in democracies and autocracies.

To explore the sensitivity of inferences to modeling choices, we replicated the main specification of Escribà-Folch, Meseguer, and Wright (2018) (Table 1, column 2), using each of the methods discussed above. Rather than interact only *Remit* with regime type, we re-estimate the interaction between *Remit* and *Autocracy* using a fully moderated model, where *Autocracy* is interacted with all covariates and fixed effects. We also estimate the same effect using the post-double-selection approach, allowing small interactions between control variables (including fixed effects) and *Autocracy* to be removed from the analysis.

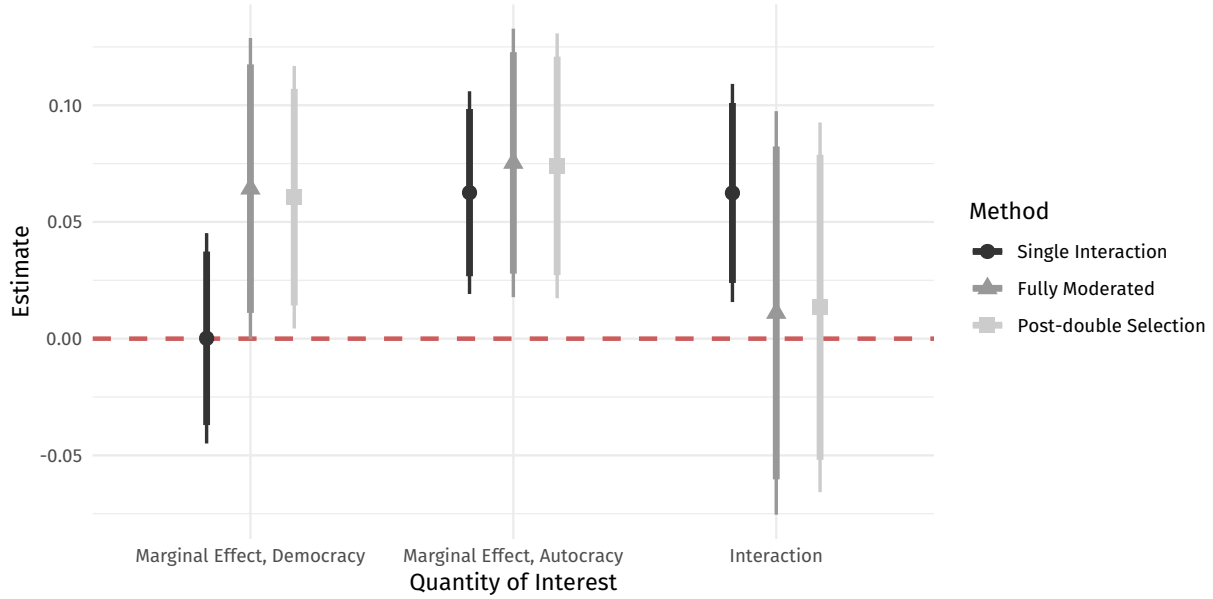

Figure 5: Remittances, Protest, and Regime Type

Estimates from interactive model originally reported in Escribà-Folch, Meseguer, and Wright (2018) and estimates from the fully moderated and post-double-selection models. 90% and 95% confidence intervals are based on regime-clustered standard errors.

Figure 5 plots points estimates and confidence intervals from these approaches. We report these estimates for three quantities of interest: the marginal effect of remittances in autocracies, the marginal

effect of remittances in democracies, and the interaction between remittances and autocracy. As Figure 5 makes clear, estimates differ considerably depending on the estimator used. All models are consistent in their conclusion that remittances are important predictors of protest in autocracies, but only the single-interaction model supports the authors' original conclusion that they are *not* related to protest in democracies. Estimates from both the fully moderated model and its regularized version suggest that remittances have an almost equally sized effect on political protest in democracies and autocracies—similar, in both cases, to the effect size originally reported by the authors for autocracies. Our estimate of the moderating effect of regime type is near-zero and statistically indistinguishable from it. As expected, the use of the post-double-selection estimator produces somewhat tighter confidence intervals than the fully moderated model, indicative of the regularized model's value in preserving statistical power by eschewing irrelevant moderator-covariate interactions.

6 Conclusion

In this paper, we highlight an overlooked issue in the estimation of interactive effects in regression models. Namely, we show how a single multiplicative interaction term can be biased when interactions between the same moderator and other covariates are omitted from the model. These omitted interaction can considerably change the estimated effect heterogeneity and lead scholars to draw misleading conclusions. To avoid this issue, we advocate for two possible solutions. The first is a fully moderated (or split-sample) model that includes an interaction between the moderator and all variables in the model. The second is a regularized version of this procedure that uses the lasso to select which of the moderator-covariate interactions are important for estimation. The latter procedure can be useful when there are large number of covariates and including all covariates can lead to imprecise estimates.

Based on our analyses, we recommend that scholars bring think carefully about model misspecification when estimating interaction, and when possible, use more flexible estimation procedures for this purpose. This includes assessing linearity of the interaction, as Hainmueller, Mummolo, and

Xu (2019) emphasize, but also to consider how lower-order terms of the moderator and covariates, along other nuisances, affect inferences. In this paper, we have focused on the lasso, but other machine learning methods may also provide flexible ways of estimating interactions. When using other machine learning methods, though, it is important to assess how they perform in terms of estimating low-dimensional parameters since many of these methods are designed for general prediction tasks and not the traditional inference of the applied social sciences.

Bibliography

Ansolabehere, Stephen, Shigeo Hirano, and Jr. Snyder, James M. 2007. "What Did the Direct Primary Do to Party Loyalty in Congress." In *Process, Party and Policy Making: Further New Perspectives on the History of Congress*, ed. David Brady and Matthew D McCubbins. Stanford University Press, Palo Alto.

Bansak, Kirk. 2018. "A Generalized Framework for the Estimation of Causal Moderation Effects with Randomized Treatments and Non-Randomized Moderators." Working paper, arXiv:1710.02954. **URL:** https://arxiv.org/abs/1710.02954

Beiser-McGrath, Janina, and Liam F. Beiser-McGrath. 2019. "Problems with Products? Control Strategies for Models with Interactive and Quadratic Effects." *Working Paper*. **URL:** https://www.dropbox.com/s/enojv35z5bv8edz/beisermcgrath_interactions_comb.pdf?raw=1

Belloni, Alexandre, Victor Chernozhukov, and Christian Hansen. 2014. "Inference on Treatment Effects after Selection among High-Dimensional Controls." *The Review of Economic Studies* 81 (11): 608–650.

Belloni, Alexandre, Victor Chernozhukov, Christian Hansen, and Damian Kozbur. 2016. "Inference in High-Dimensional Panel Models With an Application to Gun Control." *Journal of Business & Economic Statistics* 34 (4): 590-605.

Berry, William D., Jacqueline H. R. DeMeritt, and Justin Esarey. 2010. "Testing for Interaction in Binary Logit and Probit Models: Is a Product Term Essential?" *American Journal of Political Science* 54 (1): 248-266.

Bien, Jacob, Jonathan Taylor, and Robert Tibshirani. 2013. "A lasso for hierarchical interactions." *Annals of statistics* 41 (3): 1111.

Brambor, Thomas, William Roberts Clark, and Matt Golder. 2006. "Understanding Interaction Models: Improving Empirical Analyses." *Political Analysis* 14 (1): 63–82.

Braumoeller, Bear F. 2004. "Hypothesis Testing and Multiplicative Interaction Terms." *International Organization* 58 (4): 807–820.

Camponovo, Lorenzo. 2015. "On the Validity of the Pairs Bootstrap for Lasso Estimators." *Biometrika* 102 (4): 981-987.

Esarey, Justin, and Jane Lawrence Sumner. 2018. "Marginal Effects in Interaction Models: Determining and Controlling the False Positive Rate." *Comparative Political Studies* 51 (9): 1144-1176.

Escribà-Folch, Abel, Covadonga Meseguer, and Joseph Wright. 2018. "Remittances and Protest in Dictatorships." *American Journal of Political Science* 62 (4): 889-904.

Franzese, Robert J, and Cindy Kam. 2009. *Modeling and interpreting interactive hypotheses in regression analysis*. Ann Arbor: University of Michigan Press.

Hainmueller, Jens, Jonathan Mummolo, and Yiqing Xu. 2019. "How Much Should We Trust Estimates from Multiplicative Interaction Models? Simple Tools to Improve Empirical Practice." *Political Analysis* 27 (2): 163-–192.

Hirano, Shigeo, and James M Snyder. 2007. "The Decline of Third-Party Voting in the United States." *Journal of Politics* 69 (1): 1–16.

Hirano, Shigeo, and James M. Snyder, Jr. 2019. *Primary Elections in the United States*. Cambridge: Cambridge University Press.

Imai, Kosuke, and Marc Ratkovic. 2013. "Estimating treatment effect heterogeneity in randomized program evaluation." *The Annals of Applied Statistics* 7 (03): 443–470.

Imbens, Guido W. 2004. "Nonparametric Estimation of Average Treatment Effects Under Exogeneity: A Review." *Review of Economics and Statistics* 86 (February): 4–29.
URL: http://www.mitpressjournals.org/doi/abs/10.1162/003465304323023651

Kam, Cindy D, and Marc J Trussler. 2017. "At the nexus of observational and experimental research: Theory, specification, and analysis of experiments with heterogeneous treatment effects." *Political Behavior* 39 (4): 789–815.

Knight, Keith, and Wenjiang Fu. 2000. "Asymptotics for lasso-type estimators." *The Annals of Statistics* 28 (10): 1356–1378.

Kousser, J. Morgan. 1974. *The Shaping of Southern Politics: Suffrange Restrictions and the Establishment of the One-Party South, 1880-1910*. New Haven: Yale University Press.

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

Merriam, Charles Edward. 1908. *Primary elections: A study of the history and tendencies of primary election legislation*. Chicago: University of Chicago Press.

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

Perman, Michael. 2001. *Struggle for Mastery: Disfranchisement in the South, 1888-1908*. Chapel Hill: University of North Carolina Press.

Ratkovic, Marc, and Dustin Tingley. 2017. "Sparse estimation and uncertainty with application to subgroup analysis." *Political Analysis* 25 (1): 1–40.

Reynolds, John F. 2006. *The Demise of the American Convention System, 1880–1911*. Cambridge: Cambridge University Press.

Tibshirani, Robert. 1996. "Regression shrinkage and selection via the lasso." *Journal of the Royal Statistical Society. Series B (Methodological)* 58 (1): 267–288.

VanderWeele, Tyler. 2015. *Explanation in causal inference: methods for mediation and interaction*. Oxford: Oxford University Press.

Vansteelandt, Stijn, Tyler J. VanderWeele, Eric J. Tchetgen, and James M. Robins. 2008. "Multiply Robust Inference for Statistical Interactions." *Journal of the American Statistical Association* 103 (484): 1693–1704.

Ware, Alan. 2002. *The American Direct Primary*. Cambridge: Cambridge University Press.

Zou, Hui, and Trevor Hastie. 2005. "Regularization and variable selection via the elastic net." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67 (2): 301–320.

Supplemental Materials (to appear online)

A Additional results

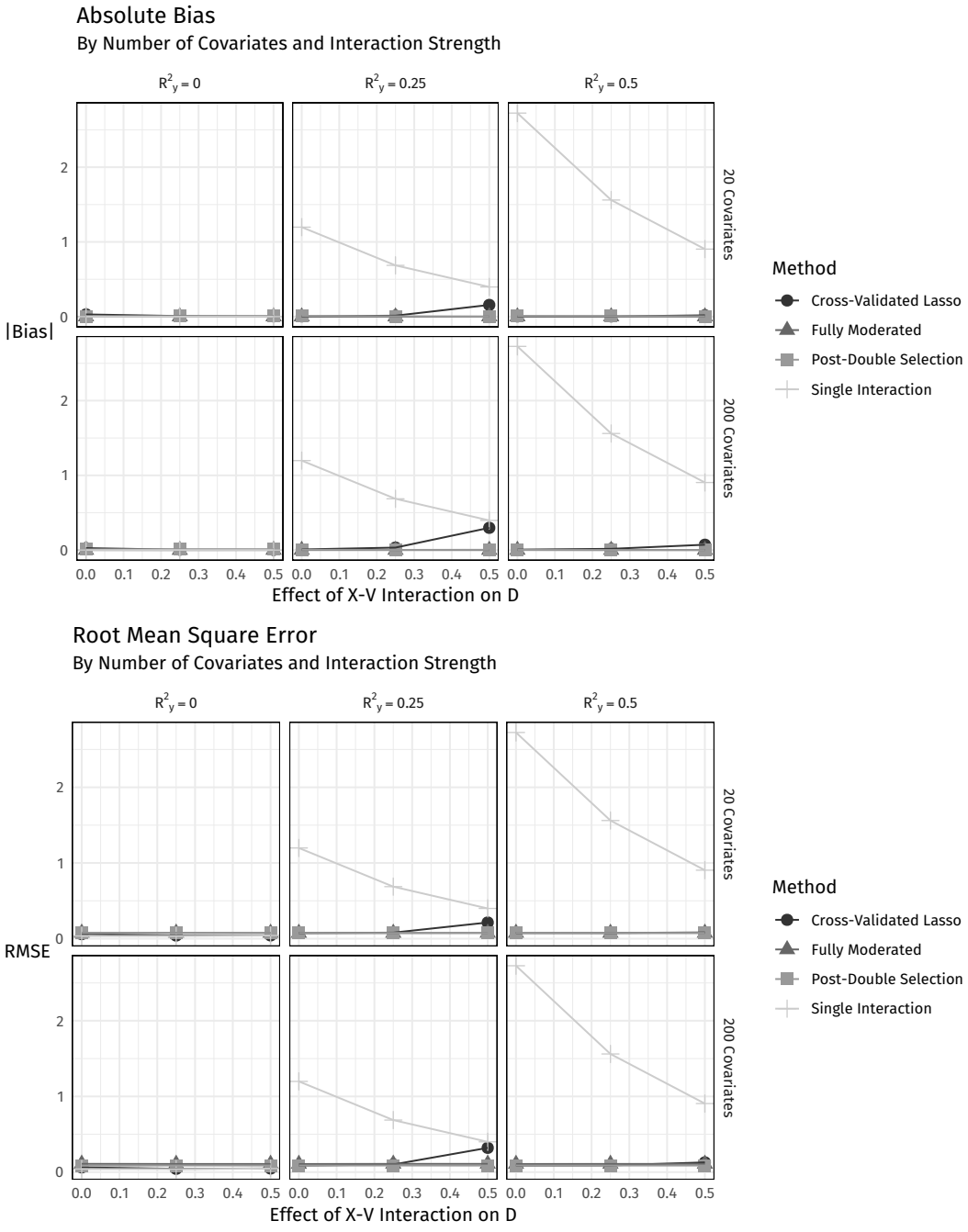

Figure SM.6: Simulation results including single interaction model