Reducing Model Misspecification and Bias in the Estimation of Interactions∗

Matthew Blackwell† and Michael Olson‡

April 10, 2020

Abstract

Analyzing variation in treatment effects across subsets of the population is an important way for social scientists to evaluate theoretical arguments. A common strategy in assessing 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 include the additional interactions, but this can lead to unstable estimates due to overfitting. Machine learning algorithms can stabilize these estimates but can also lead to bias due to regularization and model selection mistakes. To overcome these issues, we use a post-double selection approach that utilizes several lasso estimators to select the interactions to include in the final model. We extend this approach to estimate uncertainty for both interaction and marginal effects. Simulation evidence shows that this approach has lower bias and uncertainty than competing methods, even when the number of covariates is large. We show in two empirical examples that the choice of method leads to dramatically different conclusions about effect heterogeneity.

1 Introduction

The social and political worlds are full of heterogeneity. Exploring such heterogeneity in treatment effects has become an important and widely used approach 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, 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 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 interactive models. 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.

If single interaction terms can create such bias, what alternative do applied researchers have? One approach, analogous to a split-sample strategy that one might use with a discrete moderator, 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, 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. To do so, we use a variant 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 guard against large biases due to misspecification while reducing inclusion of irrelevant interactions that reduce statistical efficiency. Finally, we propose a new variance estimator for the post-double selection approach that captures the covariance between estimated coefficients, which allows for the estimation of uncertainty estimates for both the interaction and marginal effects.

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 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. In our simulation study, we find that single-selection approaches like the standard outcome lasso can have significantly higher bias compared to post-double selection.

In developing our intuitions and solutions, 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). We differ from these studies, however, in that we focus on methods for estimating a particular interaction of theoretical interest. Our goal is to offer intuition and estimators that applied researchers can readily use to execute their already-planned analyses of interactive effects, and so our recommendations are suited for “confirmatory” analyses about a small number of quantities of interest rather than “exploratory” analyses that seek to find interesting heterogeneous effects with limited theoretical guidance.

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 proposed variance estimator and our extension for handling fixed effects in this setting. 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:

\[
Y_i = \alpha_0 + \alpha_1 D_i + \alpha_2 V_i + X_i' \alpha_3 + \varepsilon_{i1} \quad (1)
\]

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 model:

\[
Y_i = \beta_0 + \beta_1 D_i + \beta_2 V_i + X_i' \beta_3 + \beta_4 D_i V_i + \varepsilon_{i2}, \quad (2)
\]

where \( \beta_4 \) 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
Figure 1: An simulated example of model misspecification in interaction models.

to running a fully moderated model, where $V_i$ is interacted with all of the variables:

$$Y_i = \delta_0 + \delta_1 D_i + \delta_2 V_i + X_i'\delta_3 + \delta_4 D_i V_i + V_i X_i'\delta_5 + \varepsilon_{i3}$$  \hspace{1cm} (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, $\hat{\beta}_4$. Under the standard omitted variable bias formula, we have $\hat{\beta}_4 \xrightarrow{p} \delta_4 + \gamma'_\nu\delta_5$, where $\gamma'_\nu$ 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, $\gamma'_\nu\delta_5$, 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.

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 and the moderator as:

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

(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 \( Y_i(d) \perp D_i \mid V_i, X_i \), where \( A \perp B \mid C \) 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

\[
\delta(d, d^*) = \int_{x \in \mathcal{X}} (\mathbb{E}[Y_i \mid D_i = 1, V_i = 1, X_i = x] - \mathbb{E}[Y_i \mid D_i = 0, V_i = 1, X_i = x]) dF_{X|V}(x|V_i = 1)
- \int_{x \in \mathcal{X}} (\mathbb{E}[Y_i \mid D_i = 1, V_i = 0, X_i = x] - \mathbb{E}[Y_i \mid D_i = 0, V_i = 0, X_i = x]) dF_{X|V}(x|V_i = 0),
\]

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 \( \delta(d, d^*) = \beta \times (d - d^*) \), and in the fully moderated model we have \( \delta(d, d^*) = \beta \times (d - d^*) \). 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 \textit{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 \textit{average} or \textit{overall} effect of treat-
ment. 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
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.

As a solution to these concerns, we propose using a standard approach to guarding against overfitting: regularization. Specifically, we perform variable selection and estimation of model parameters through use of the 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 asymptotic bias in effect estimates for retained coefficients, even when the number of covariates is fixed (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 \( D_iV_i \)) 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. 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 these 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 \( D_iV_i \) as the main variables of interest and let \( Z^*_i = [V_i \ X'_i \ V_iX'_i] \) 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 \( \{Y_i, D_i, D_iV_i \} \) 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).

\[
\hat{\gamma}_y = \arg \min_{\gamma_y} \sum_{i=1}^{N} (Y_i - Z^*_i\gamma_y)^2 + \sum_{j=1}^{k} \lambda_{yj}|\gamma_{yj}| \\
\hat{\gamma}_d = \arg \min_{\gamma_d} \sum_{i=1}^{N} (D_i - Z^*_i\gamma_d)^2 + \sum_{j=1}^{k} \lambda_{dj}|\gamma_{dj}| \\
\hat{\gamma}_{d^v} = \arg \min_{\gamma_{d^v}} \sum_{i=1}^{N} (D_iV_i - Z^*_i\gamma_{d^v})^2 + \sum_{j=1}^{k} \lambda_{d^v j}|\gamma_{d^v j}|
\]

Let \( Z^*_i \) be a vector of the subset of \( Z_i \) that has either \( \hat{\gamma}_y, \hat{\gamma}_d, \) or \( \hat{\gamma}_{d^v} \) not equal to zero. The final step of post-double selection is to regress \( Y_i \) on \( \{D_i, D_iV_i, Z^*_i \} \) 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 func-
tions 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.\footnote{For example, let $Z_i'\gamma_{y0}$ be a sparse approximation to the outcome CEF in that the number of non-zero values in $\gamma_{y0}$ is less than some fixed values $s$. Define the approximation error $r_i = \mathbb{E}[Y_i \mid Z_i] - Z_i'\gamma_{y0}$. Then, a CEF is approximately sparse if $(\mathbb{E}[N^{-1} \sum_i r_i^2])^{1/2} \leq C \sqrt{s/N}$ as $N \to \infty$.} 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 asymptotic results of Belloni, Chernozhukov, and Hansen (2014) are valid for high-dimensional models, where the number of covariates or parameters in the model grows with the sample size. Our discussion, on the other hand, has focused on a model where the number of covariates is fixed but could be large once all $X_iV_j$ interactions are added to the model. The usual asymptotic results for fixed-parameter models would imply that the fully moderated model should outperform the post-double selection approach, but asymptotic results are only useful insofar as they predict performance in finite, realistic sample sizes which we investigate in the simulations below. Furthermore, when the number of covariates is large relative the sample size, the fully moderated model will become either unstable or not possible to calculate, but PDS should maintain its desirable properties.

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, $\hat{\varepsilon}_{yi}$, as with robust variance estimators. Belloni, Chernozhukov, and Hansen (2014) show that this procedure (along with a carefully chosen value of the $\lambda_{y0}$) ensures consistency and asymptotic normality even when the errors are non-normal and 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.

### 3.3 Variance estimator for interactions after post-double selection

In the interaction setting, we are often interested in making inferences on both the interaction term itself and various marginal effects of the main treatment. This task requires joint inference for all parameters and, in particular, the covariance between the lower-order and interaction terms. The original post-double selection approach of Belloni, Chernozhukov, and Hansen (2014) handled multiple parameters of interest by applying the approach for a single parameter to each variable of interest separately, which does not allow for this type of joint inference.\(^2\)

We propose an alternative variance estimator that also estimates the covariance between the estimated effects of \( D_i \) and \( D_i V_i \) in order to quantify uncertainty for marginal effects. In particular, we define \( \tilde{Y}_i, \tilde{D}_i, \) and \( \tilde{DV}_i \) to be the residuals from regressions of \( Y_i, D_i, \) and \( D_i V_i \) on \( Z_{i}^{*} \) (the selected set

\(^2\)Belloni, Chernozhukov, and Kato (2014) does propose a bootstrap method for generating uniformly-valid joint confidence regions for multiple parameters. This, however, does not help the typical use case with interactions in the social sciences, where we are interested in confidence intervals for the marginal effects which are linear functions of the estimates.
of covariates from the double selection). Let \( \tilde{\delta}_i = \tilde{Y}_i - \tilde{\delta}_1 \tilde{D}_i - \tilde{\delta}_4 \), where \( \tilde{\delta}_1 \) and \( \tilde{\delta}_4 \) are the post-double selection estimators of the coefficients on \( D_i \) and \( D_i V_i \), respectively. Let \( D \) be the matrix with rows \( (\tilde{D}_i, \tilde{D}_i V_i)^T \) and define the following projection matrix: \( H = D(D'D)^{-1}D' \). Let \( h_i = h_{ii} \) be the diagonal entries of this matrix. Then, we define \( \hat{\Omega} \) to be a diagonal matrix with entries \( \tilde{\delta}_i^2 / (1 - h_i)^2 \).

Then, our estimated covariance matrix of \( (\hat{\delta}_1, \hat{\delta}_4) \) is has the following sandwich form:

\[
\tilde{V} = \frac{1}{N} \times \frac{N - 1}{N - K^* - 3} (D'D)^{-1} \left( D' \hat{\Omega} D \right) (D'D)^{-1},
\]

where \( K^* \) is the dimension of \( Z_i^\star \). Essentially, this is a heteroskedastic-consistent variance estimator of MacKinnon and White (1985) applied to the residualized regression. This generalizes the univariate version of this estimator that Belloni, Chernozhukov, and Hansen (2014) applied to each coefficient separately. Below, we show that this estimator produces confidence intervals with good coverage under the approximate sparsity setting that Belloni, Chernozhukov, and Hansen (2014) investigated. While these covariances are important for the interaction setting, this approach would be useful anytime a researcher is interested in a function of the parameters of interest.

### 3.4 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 significant 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

<table>
<thead>
<tr>
<th></th>
<th>(R_{i1})</th>
<th>(R_{i2})</th>
<th>(R_{i3})</th>
</tr>
</thead>
<tbody>
<tr>
<td>Northeast</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
</tr>
<tr>
<td>Midwest</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>West</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>South</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 1: Deviation coding example
\( N = \sum_{g=1}^{G} N_g \) total individuals. Then, we would set the penalty parameter as \( \lambda_{yj} = \lambda_{y0} \phi_{yj} \), where

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

For a feasible estimate of this penalty, we can run an initial lasso to obtain estimates of \( \hat{\varepsilon}_{yig} \). The penalty terms for the other lasso regressions follow similarly. Again, the penalty parameter depends on a measure of the noise in estimating the \( \gamma_{yj} \), 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. Finally, we can easily extend the above variance estimator to handle clustering by changing the form of the above estimator to that of a cluster-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 \( X_t \) of dimension \( K \), from \( \mathcal{N}(0, \Sigma) \), where \( \Sigma_{ij} = 0.5|j-k| \) so that the covariates depend on each other. We set the sample size to 425 and 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 as

\[
\mathbb{P}[V_i = 1 \mid X_i] = \log^{-1}(\delta_{v[0]} + X_i' \delta_{v|x}),
\]

with the treatment and outcome as:

\[
D_i = \delta_{d[0]} + 0.25 \times V_i + X_i' \delta_{d|x} + V_i X_i \delta_{d|ux} + \varepsilon_{id}
\]

\[
Y_i = \delta_{y[0]} + 0.5 \times D_i + 0.25 \times V_i + X_i' \delta_{y|x} + 1 \times D_i V_i + V_i X_i \delta_{y|ux} + \varepsilon_{iy}
\]

The parameters of these models are generated under a quadratic decay, so that the \( j \)th entry of \( \delta_{v|x} \) is \( \delta_{v|x[j]} = 2/j^2 \). We define the other coefficient vectors similarly: \( \delta_{d|x[j]} = 2/j^2 \), \( \delta_{y|x[j]} = 2/j^2 \).
\[ \delta_{d|x_{ij}} = \frac{c_{d|x}}{j^2}, \quad \text{and} \quad \delta_{y|x_{ij}} = \frac{c_{y|x}}{j^2}. \]

We vary \( c_{d|x} \) and \( c_{y|x} \) so that the \( V_i X_i \) interactions have partial \( R^2 \) values of \( \{0, 0.25, 0.5\} \). Each of the errors, \( \{\xi_{id}, \xi_{iy}\} \), 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-single selection approach that only selects variables based on a lasso applied to the outcome, using the data-driven penalty choices of Belloni and Chernozhukov (2013). The second lasso-based procedure we apply is the above post-double-selection estimator with the same approach for choosing the penalty terms. In both approaches, we force the lower-order terms to be included in the post-selection models, so any differences between the methods are only due to their estimation of the interactions. Finally, for reference, we also estimate an infeasible “oracle” model, where we assume \( \delta_{y|0}, \delta_{y|x}, \) and \( \delta_{y|x} \) are known. For both lasso approaches we use the variance estimator described in Section 3.3, and for the other models we use the equivalent heteroskedastic variance estimator.

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.7. With a low number of covariates (\( K = 20 \)), the fully moderated model dominates the feasible methods in terms of bias, across all settings. The performance of the two lasso methods depends more on the strength of the interactions. When the interactions are unrelated to the outcome, the bias of both lasso approaches is fairly low and the post-single selection lasso actually outperforms all feasible methods in terms of RMSE. As the strength of the interactions for either \( D_i \) or \( Y_i \) increases, we see both the bias and RMSE of post-single selection lasso get larger, where the differential between them is largest when the interactions matter more for the treatment than they do for the outcome. This is what we would expect since these are the situations where we expect the single selection approach to make selection errors. When the interactions are important for both \( Y_i \) and \( D_i \), both approaches are likely to select the same variables.

In the high-dimensional setting (\( K = 200 \)), the fully moderated model is very numerically unsta-
Bias (top) and root mean square error (bottom) of various methods when estimating interactions. Horizontal panels vary the partial $R^2$ of the $V_i X_j$ 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 $V_i X_j$ on $D_j$.

Figure 2: Simulation results
ble since the number of parameters (403) is close to the sample size (425) leading to RMSE that is too high to show on the graphs so we omit it. The results on the lasso methods are remarkably similar here to in the low-dimensional setting, with slightly higher bias and RMSE across both methods. But post-double selection still outperforms single selection when the interactions are more important for the treatment than the outcome.

The results on the lasso methods are remarkably similar here to in the low-dimensional setting, with slightly higher bias and RMSE across both methods. But post-double selection still outperforms single selection when the interactions are more important for the treatment than the outcome.

In Figure 3 we present the empirical coverage of nominal 95% confidence intervals from these estimators. With a small number of covariates, both the fully moderated and post-double selection approaches perform well, with post-double selection having coverage slightly closer to nominal levels except when the interactions are strongly related to the outcome, when it slightly undercovers. The confidence intervals from the outcome-only lasso undercovers quite severely across a range of settings. With a high number of covariates, the post-double selection approach maintains its roughly

3For instance, the variance estimators for the OLS are not obtainable in fully moderated model with $K = 200$ and the RMSE of the estimator itself is roughly 100 times the worst performance of post-double selection. In simulations not shown here, we find that post-double selection can improve RMSE over the fully moderated model when the sample size is higher and thus the fully moderated model is estimable.
nominal coverage, whereas the post-single selection shows an exaggerated pattern of its performance in the low-dimensional setting. Thus, in this quadratic decay setting, where the interactions are approximately sparse, the post-double selection estimator performs well in low- and high-dimensional settings, even when fully moderated models are infeasible.

![Root Mean Square Error](image)

**Figure 4: Simulation results under a dense coefficient setting**

Root mean square error of various methods when estimating interactions. Horizontal panels vary the partial $R^2$ of the $V_iX_j$ 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 $V_iX_j$ on $D_i$.

When can the lasso approaches to this problem fail? We investigate this with an alternative data-generating process where the covariate effects are more “dense.” In particular, we set the $\delta$ effects defined above to be functions of $j^{-1}$ instead of $j^{-2}$, which spreads the same explanatory power over a larger set of covariates. We present the RMSE of the various estimators in Figure 4, where it is clear that the lasso-based estimators perform much worse than in the approximately sparse setting. For example, when $R^2_y = 0.75$, the post-double selection approach had an RMSE that is roughly ten times higher in the dense setting than in the quadratic decay setting, and the post-single selection approach has RMSE close to that of the single-interaction model. It is interesting to note, however, that post-
double selection still outperforms post-single selection and the single Interaction approaches in this setting. Furthermore, in high-dimensional settings, these lasso estimators may be the only options available to applied researchers.

Overall, the lasso-based methods can help improve bias over single-interaction methods and can improve efficiency (and estimability) over fully moderated models. The post-double selection approach appears to outperform the post-single selection approach unless interactions are unimportant for either the treatment or the outcome, an unrealistic setting in our view. Finally, the post-double selection approach appears to work best when the covariate interactions are either sparse or approximately sparse.

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 confounders. Yet in the South the direct primary was a fundamentally different reform, tied up in the disfranchisement of African Americans and the consolidation of white Democratic one-party rule (Ware, 2002, 18-20). With varying motivations for primary adoption across North and South, it is important to evaluate whether the effect of direct primary adoption is similar in the two regions.

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.4

We measure direct primary adoption as an indicator variable for whether the direct primary was in widespread use in a given state and year. Our moderator, South, is an indicator for whether a state is one of the eleven states of the former Confederacy. The single interaction model can therefore be expressed as the following:

\[
(100 - \text{DemShare}_{it} - \text{RepShare}_{it}) = \beta (\text{Primary}_{it} \times \text{South}_i) + \gamma \text{Primary}_{it} + \alpha_i + \tau_t + \varepsilon_{it} \tag{11}
\]

where \(i\) indexes states and \(t\) indexes election years. The base term on South is absorbed by the state fixed effects \(\alpha\); \(\tau\) 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 5: Effect of the Direct Primary in the American North and South

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

Figure 5 displays estimates from a single-interaction model given by Equation (11), a fully moderated model that adds interactions between the year fixed effects and the South indicator, a post-lasso estimator that adds only those year fixed effect-South interactions selected in a standard lasso procedure focused only on the outcome, and the post-double selection approach that uses a lasso to select

\[\text{We draw this information from Hirano and Snyder (2019), Table 2.A.}\]
over the interactions between the year fixed effects and the South indicator interactions for both the outcome and the direct primary “treatment.” The estimates from the single interaction and post-lasso estimates are extremely similar, and are starkly different from the fully moderated models. While all four model types agree that there is a small, statistically insignificant effect of direct primary adoption in northern states, the single-interaction and post-lasso estimators suggest a quite large effect of primary adoption 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 and post-lasso models suggest a substantially larger negative effect in the South than the North. The conclusions of the post-double-selection approach are in-between, with a marginally significant negative marginal effect of primary adoption in the South. This replication suggests key features of these different estimators. First, the post-lasso estimator, which chooses additional interactions to include based only on a lasso estimator of the outcome on covariates, here fails to select potentially impactful interactions that condition the relationship between primary adoption and region. This likely indicates that the post-lasso estimator misses that southern states tended to adopt primary elections far earlier than northern states. The replication also indicates the potential value of the post-double selection estimator over the fully moderated estimator; we see both a tightening of confidence intervals as well as possible correction to overfitting with the former relative to the latter.

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.\footnote{The authors also test their results using an instrumental variables design; we restrict our focus to their main OLS specification.} 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:

\begin{equation}
Protest_{it} = \beta (\text{Remit}_{it} \times \text{Autocracy}_{it}) + \gamma \text{Remit}_{it} + \phi \text{Autocracy}_{it} + \psi ' \mathbf{X}_{it} + \alpha_i + \tau_t + \epsilon_{it},
\end{equation}

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.

Figure 6 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 6 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 remittances matter differently, to a statistically significant degree, in autocracies and democracies. For the single interaction model, the estimated marginal effect in democracies is almost exactly zero; the interaction between remittances and autocracy is positive and significant. The post-lasso estimator, though it estimates a substantively large estimate of the marginal effect of remittances in democracies, comes closest to the single interaction model in estimating a sizable difference in the effect of remittances across regime type. Though this estimate is not statistically significant at conventional levels, it is nearly so. The fully moderated and post-double selection models, however, agree that there is little
Estimates from interactive model originally reported in Escribà-Folch, Meseguer, and Wright (2018) and estimates from the fully moderated, post-lasso, and post-double-selection models. 90% and 95% confidence intervals are based on regime-clustered standard errors.

if any difference in the effect of remittances across regime type. Interestingly, these models disagree about the extent to which remittances matter at all; the fully moderated model suggests they matter substantially in both democracies and autocracies, while the post-double selection estimates are considerably lower for both regime types, and only statistically significant in autocracies. 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. However, as we have shown, it can be important to consider both the outcome and the main independent variables when using the lasso for model selection.

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


URL: https://www.dropbox.com/s/enojv35z5bv8edz/beisermcgrath_interactions_comb.pdf?raw=1


Supplemental Materials (to appear online)

A Additional simulation results

In the main text, we omitted the simulation results for the single interaction model, which we present here in Figure SM.7. We also present the full set of bias and coverage results for the dense data-generating process in Figure SM.8.

A.1 Binary DGP

We now present results on an alternative DGP with a binary outcome. The setup is the same except for two modifications. First, we generate the outcome as a Bernoulli random variable with the following specification:

\[
\tilde{Y}_i = \delta_{y|x} + 0.25 \times V_i + X_i^T \delta_{y|x} + 1 \times D_i V_i + V_i X_i \delta_{y|x} + \varepsilon_i
\]

where \( \varepsilon_i \) follows the standard logistic distribution. Second, we modify the coefficients for the DGP. Following Belloni, Chernozhukov, and Wei (2016), we define the following:

\[
b_y = [1, 1/2, 1/3, 1/4, 1/5, 0, 0, 0, 0, 0, 1, 1/2, 1/3, 1/5, 0, 0, \ldots]^T
\]

\[
b_d = [1, 1/2, 1/3, 1/4, 1/5, 1/6, 1/7, 1/8, 1/9, 1/10, 0, 0, \ldots]^T,
\]

where the 0 values continue to make both vectors the length of the \( X_i \) (which we again vary between 20 and 200). Then, we set the coefficient values \( \delta_{d|x} = b_d/2 \), \( \delta_{y|x} = b_y/2 \), and \( \delta_{y|x} = b_d/2 \). Finally, we set \( \delta_{y|x} = c_{d|x} b_d \) and \( \delta_{y|x} = c_{y|x} b_y \). The value \( c_{d|x} \) is chosen as in the main specification to have the \( X_i V_i \) terms have a partial \( R^2 \) of 0, 0.25, 0.5. The value \( c_{y|x} \) is set so the partial \( R^2 \) of \( X_i V_i \) for the latent outcome, \( \tilde{Y}_i \) is \{0, 0.25, 0.5\}.

To apply the post-single and post-double selection methods, we use the generalized linear model setup for the lasso developed in Belloni, Chernozhukov, and Wei (2016). This setup is fairly similar to the linear modeling setup in the main text, except that the initial \( \ell_1 \)-regularized logistic regression fit for the outcome is used to produce weights for the lasso regressions for \( D_i \) and \( D_i V_i \).
selection in this case simply skips the second step. We increase the sample size to 750 to avoid numerical issues with convergence, but even in this case, the fully moderated model fails to converge when \( K = 200 \), so we omit it. The oracle model in this case selects the 15 relevant variables out of 20 or 200 to include in the outcome logistic regression.
Figure SM.8: Simulation results for bias and coverage for the dense data-generating process

Figure SM.9 displays the results. The single-selection lasso has higher bias than the post-double selection approach, even sometimes having higher bias than simply using the single-interaction model. Interestingly, this is offset by smaller variance which means all of the non-single-interaction methods have similar RMSE. The bias does have a pernicious effect on the coverage rates for the post-single
selection method, however, and they have 0 coverage. Post-double selection, on the other hand, maintains fairly good coverage across the difference specifications. Overall, this points to post-double selection being useful for estimating interactions even with binary outcomes and logistic regressions.
Figure SM.9: Simulation results for bias and coverage for the binary data-generating process.