Supplemental Materials for "Reducing Model Misspecification and Bias in the Estimation of Interactions"*

Matthew Blackwell† and Michael Olson‡

February 8, 2021

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.1. We also present the full set of bias and coverage results for the dense data-generating process in Figure SM.2.

*To be published online.

†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.

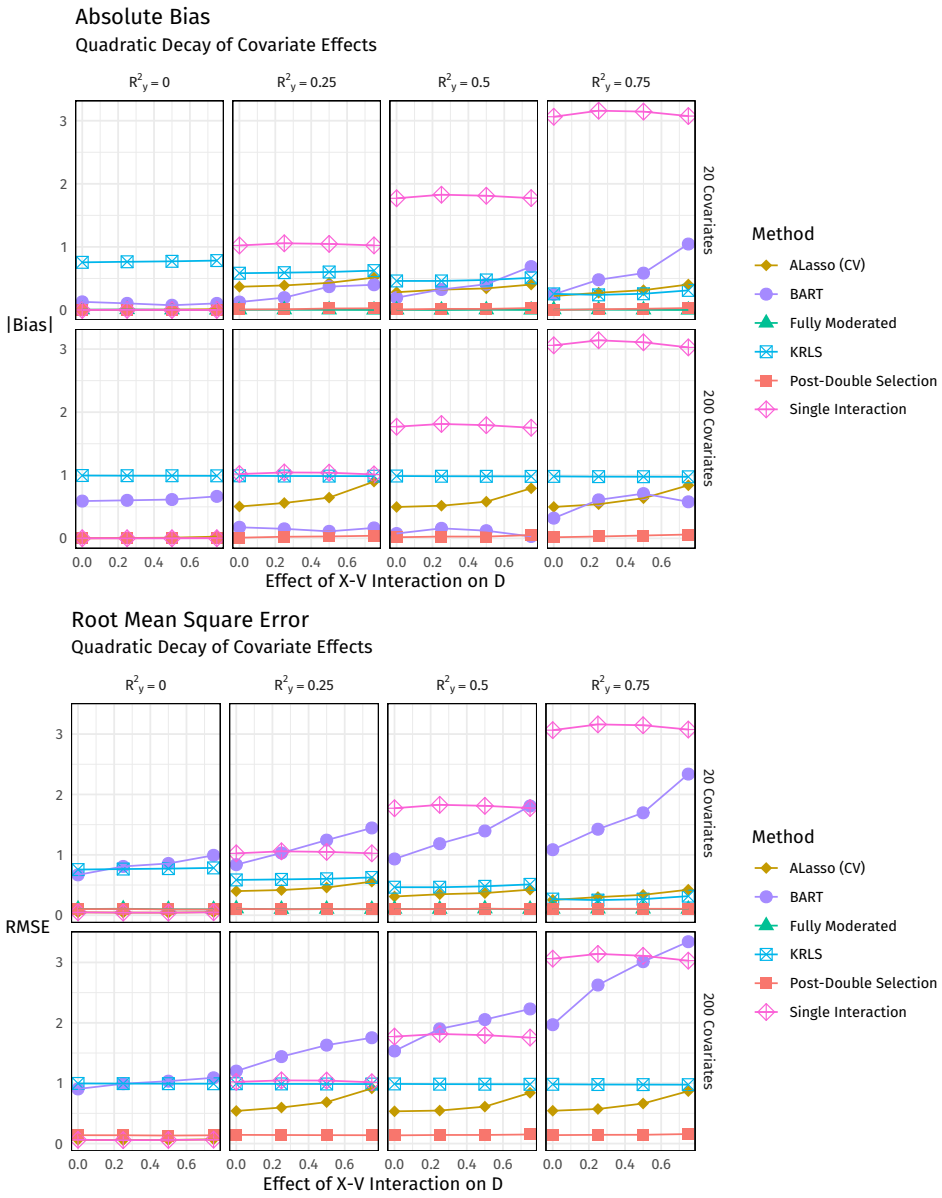

Figure SM.1: Simulation results including single interaction model

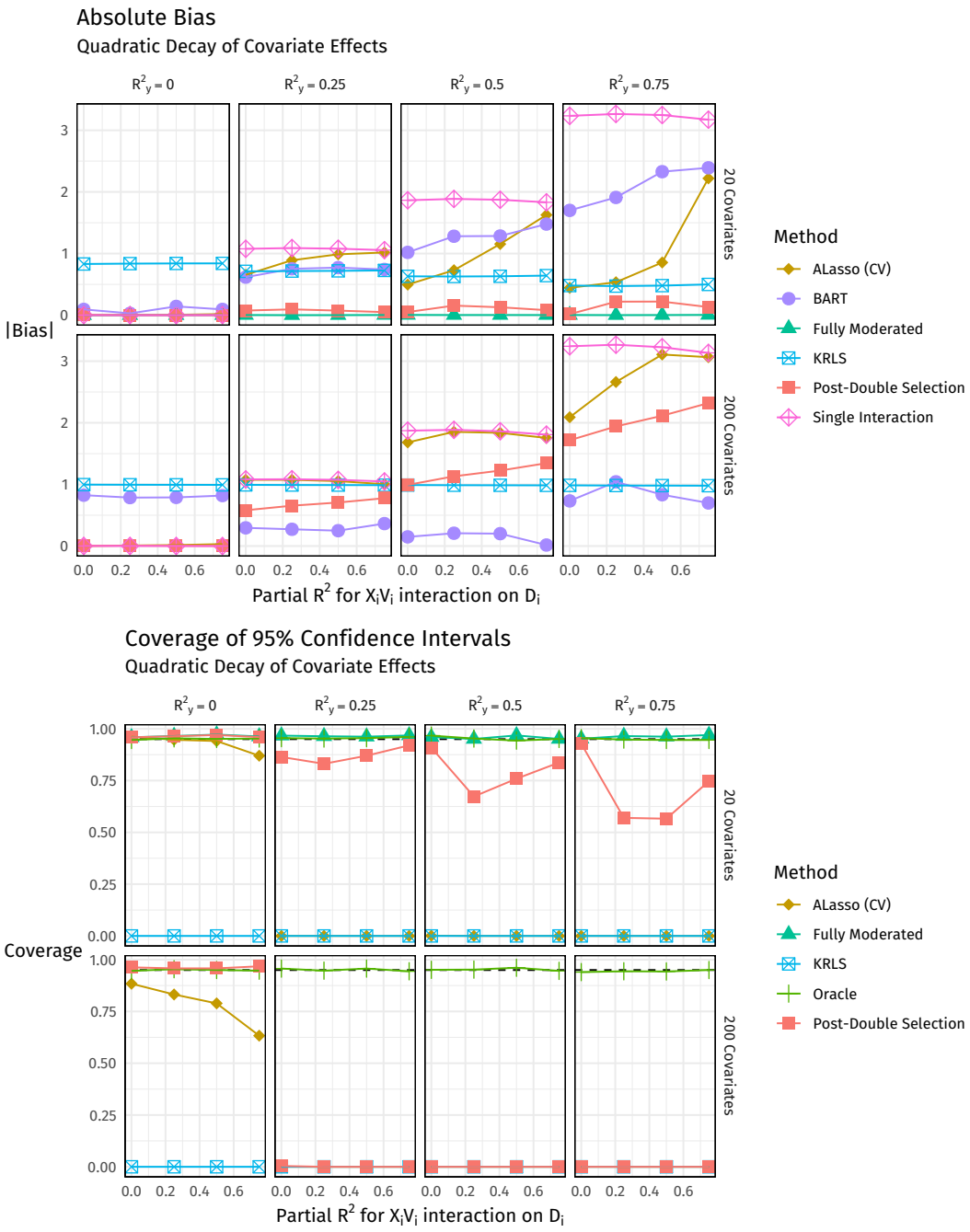

Figure SM.2: Simulation results for bias and coverage for the dense data-generating process

A.1 Post-Single Selection Lasso

To isolate the effect of direct versus indirect regularization bias, we also conducted a simulation with a post-lasso estimator. This estimator was the same as the PDS estimator in the main paper, but the covariate selection comes from the lasso applied to the outcome model only, rather than for D_i and D_iV_i as well. The results are shown in Figure SM.3 and Figure SM.4, where it seems that the post-lasso approach removes most of the direct regularization bias, but indirect regularization bias exists when the the D_i-X_iV_i relationship is strong and the Y_i-X_iV_i is moderate. This makes sense because this is the situation where the post-lasso is likely to make covariate selection mistakes that the post-double selection approach avoids.

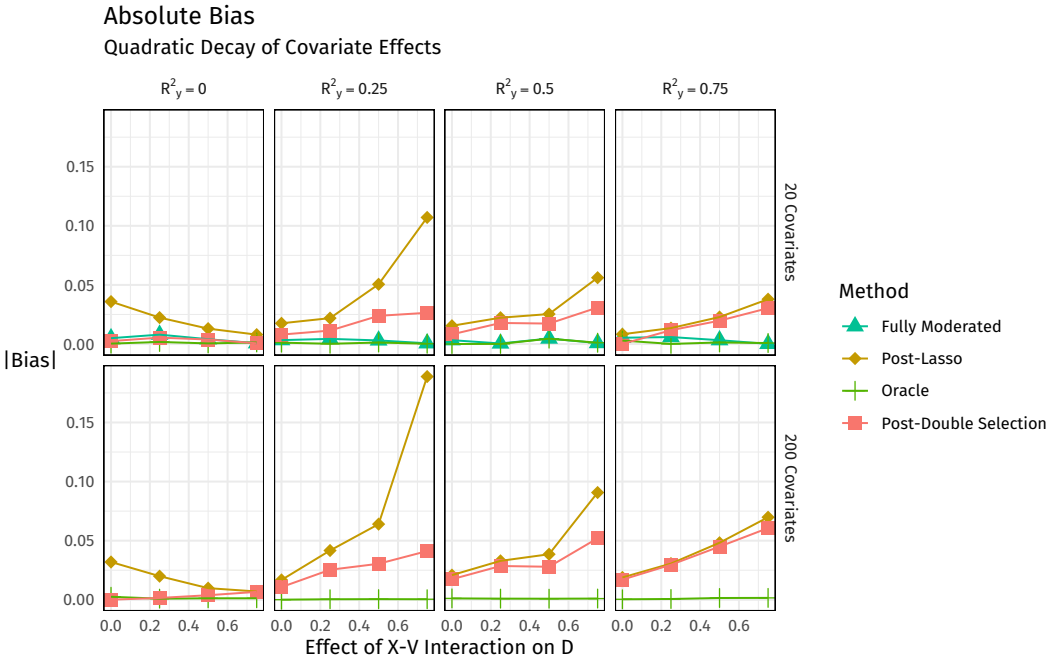

Figure SM.3: Simulation results for bias comparing post-double selection to post-lasso

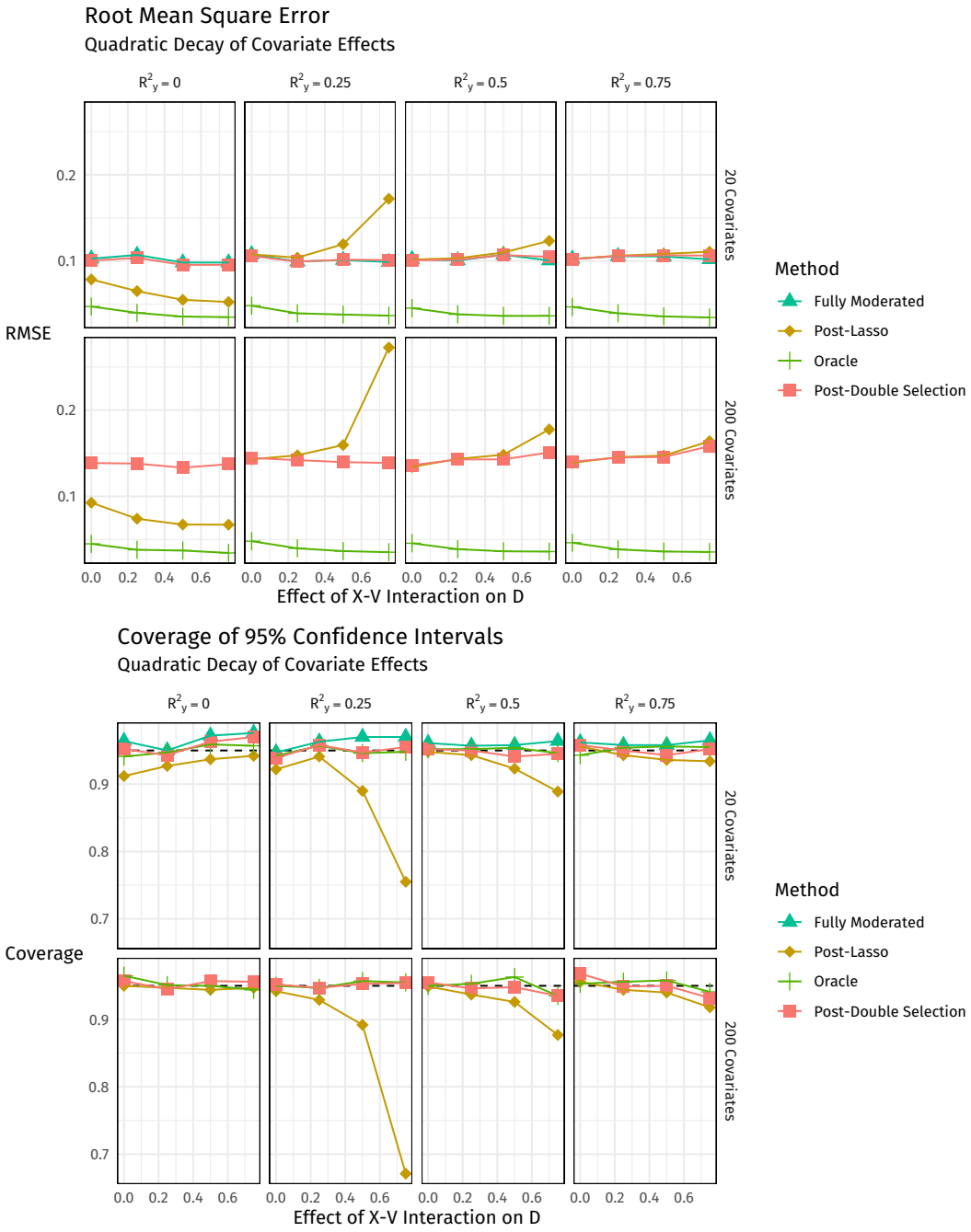

Figure SM.4: Simulation results for bias and coverage comparing post-double selection to post-lasso

A.2 Larger Sample Size

We also conducted a simulation with a larger sample size of 1000. In Figure SM.5 we show the results for the main methods discussed in the paper. Overall, the results are very similar to the sample size of 425 with some improvements in performance by the adaptive lasso. PDS and the fully moderated models continue to outperform all other methods. In Figure SM.6, we focus on the fully moderated, oracle, PDS, and post-lasso methods. Here we can see that when the number of covariates is large, PDS can outperform the fully moderated model in terms of RMSE even when the latter is feasible (unlike the $N = 425$ case). Furthermore, the fully moderated model has poor coverage performance of its confidence intervals when as the number of covariates increase.

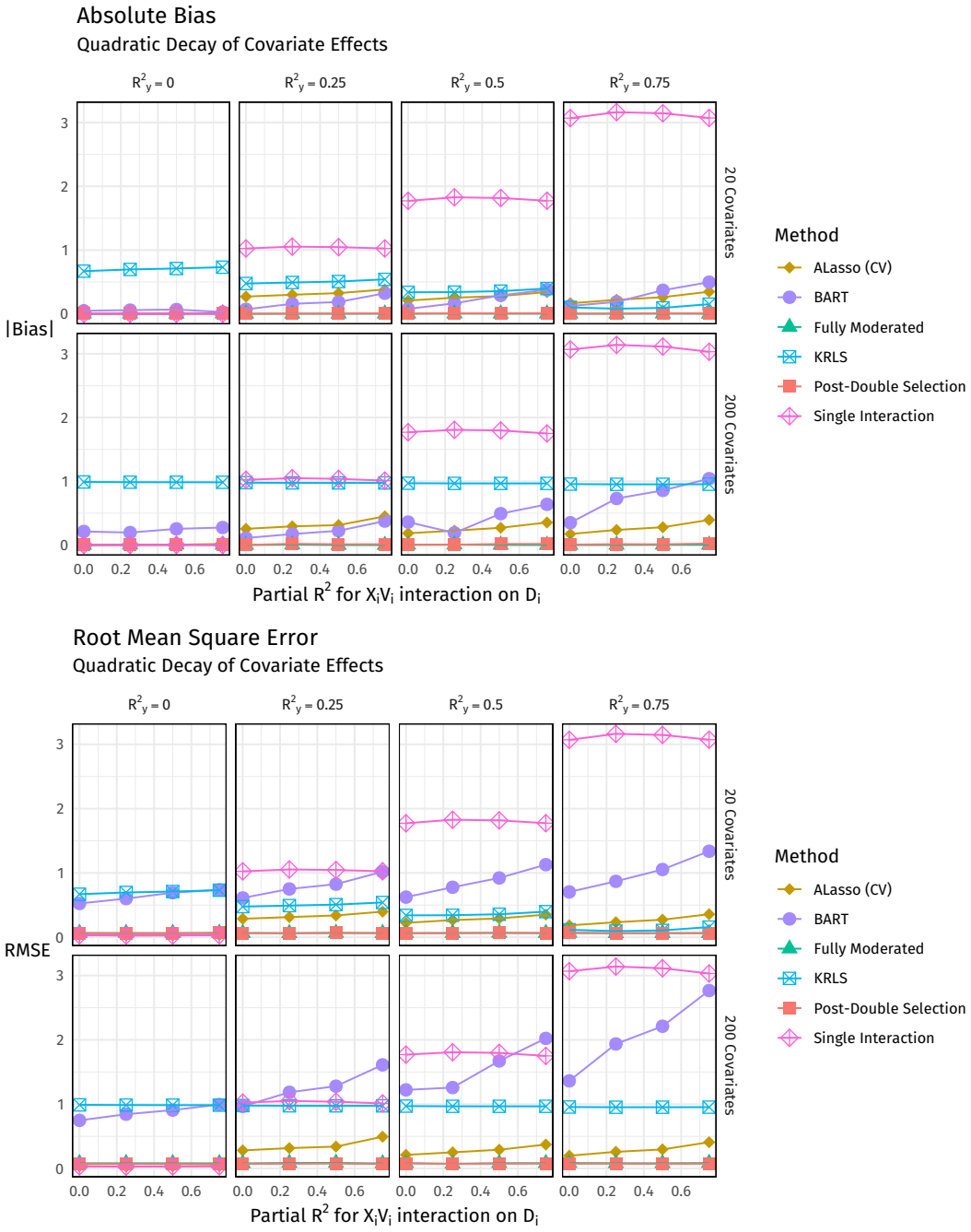

Figure SM.5: Simulation results for bias and RMSE with $N = 1000$

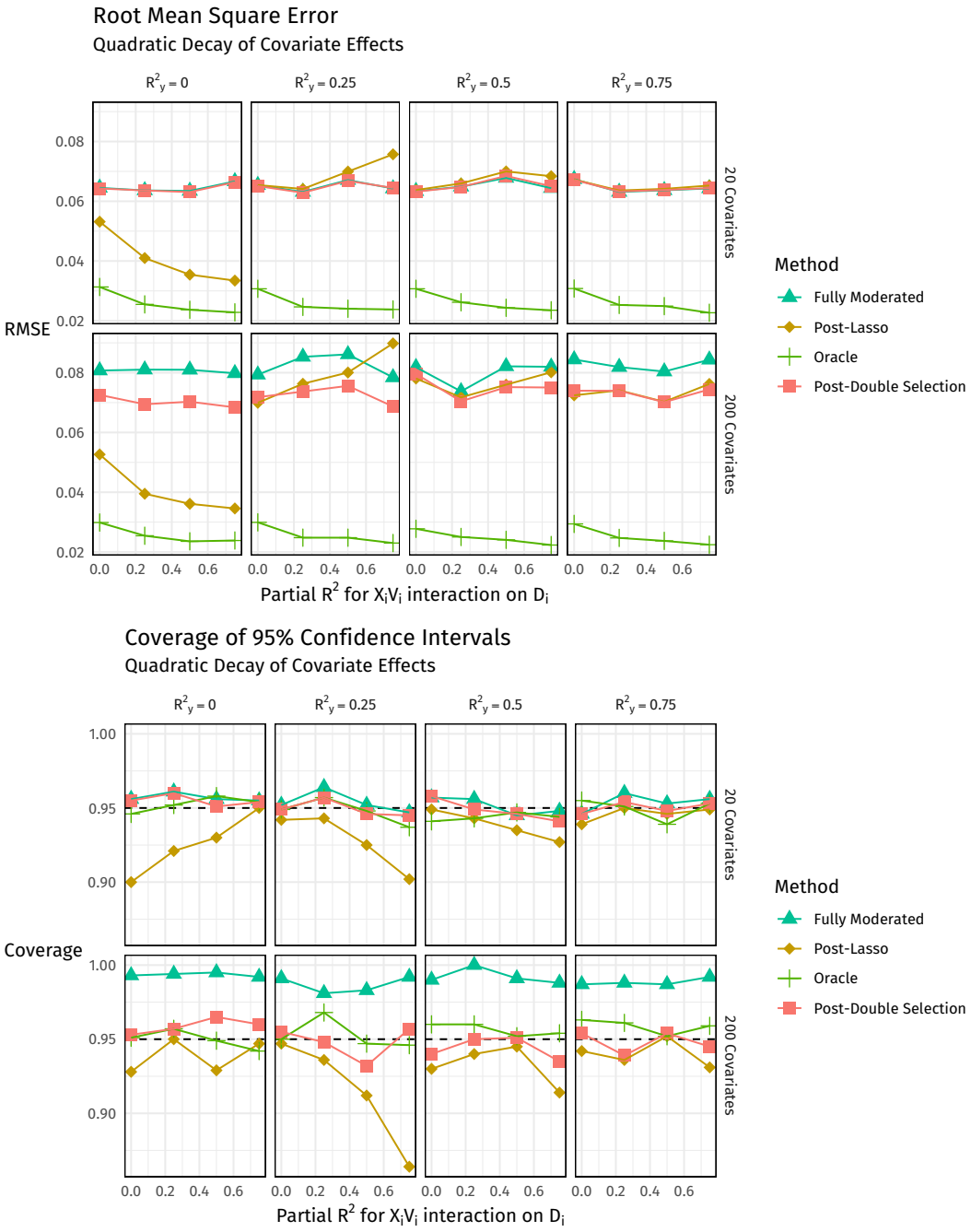

Figure SM.6: Simulation results for RMSE and coverage with $N = 1000$ comparing to the post-single selection.

A.3 Binary DGP

We now present results on a 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:

$$\widetilde{Y}_i = \delta_{y|0} + 0.1 \times D_i + 0.25 \times V_i + X_i' \boldsymbol{\delta}_{y|x} + 1 \times D_i V_i + V_i X_b \boldsymbol{\delta}_{y|vx} + \varepsilon_i$$

$$Y_i = \mathbb{I}\{\widetilde{Y}_i > 0\},$$

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

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

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

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 $\boldsymbol{\delta}_{d|x} = b_d/2$, $\boldsymbol{\delta}_{y|x} = b_y/2$, and $\boldsymbol{\delta}_{v|x} = b_d/2$. Finally, we set $\boldsymbol{\delta}_{d|vx} = c_{d|vx} b_d$ and $\boldsymbol{\delta}_{y|vx} = c_{y|vx} b_y$. The value $c_{d|vx}$ 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 $c_{y|vx}$ is set so the partial R^2 of $X_i V_i$ for the latent outcome, $\widetilde{Y}_i$ 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 ℓ_1-regularized logistic regression fit for the outcome is used to produce weights for the lasso regressions for D_i and $D_i V_i$. Post-single 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.

Figures SM.7 and SM.8 display 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.

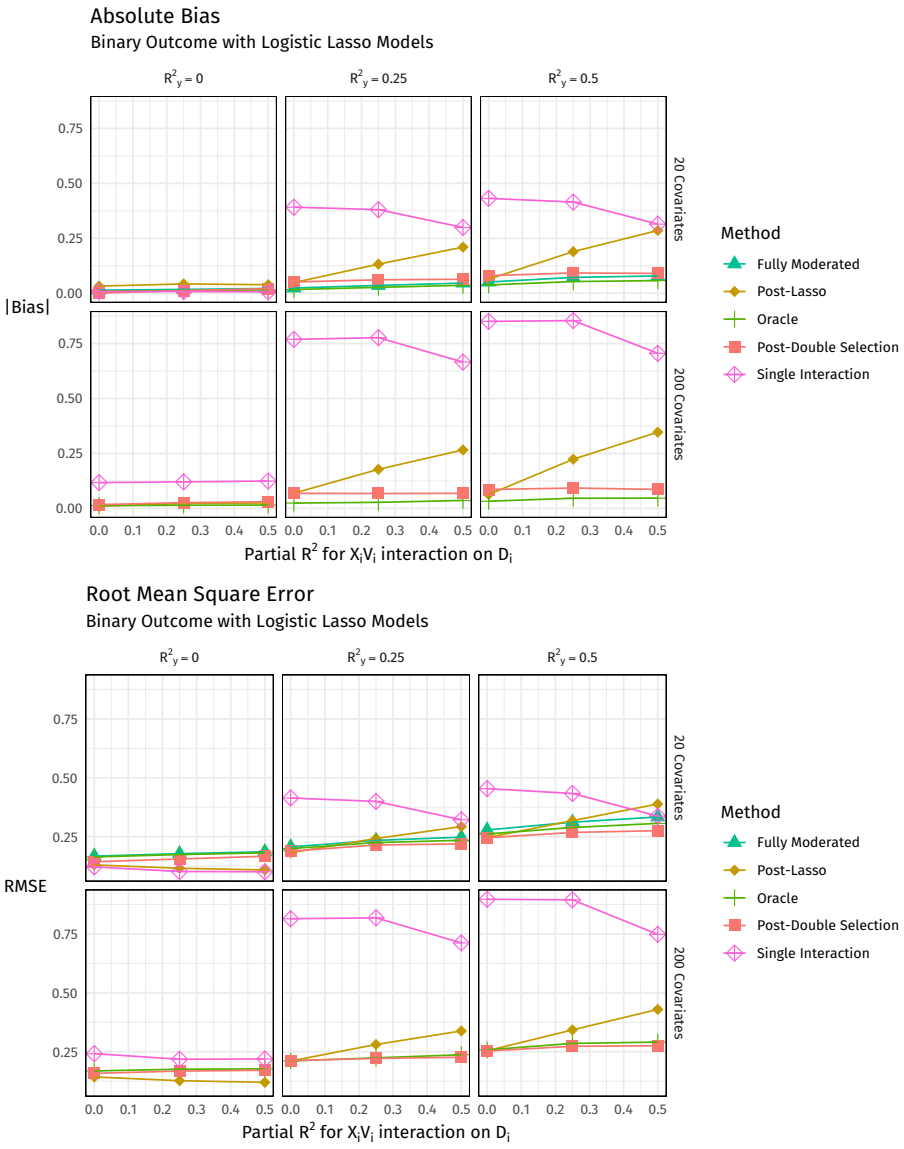

Figure SM.7: Simulation results for bias for the binary data-generating process

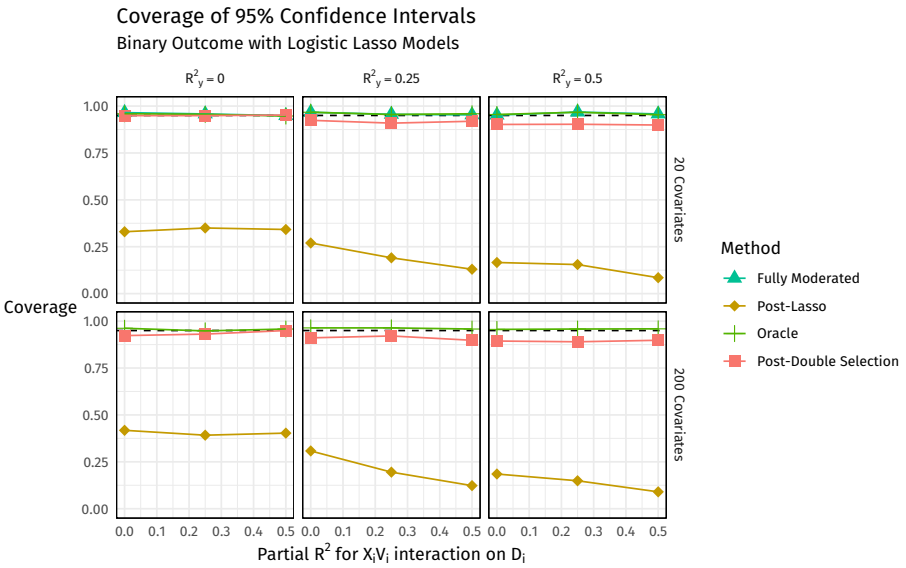

Figure SM.8: Simulation results for bias for the binary data-generating process

B Additional Replication Results

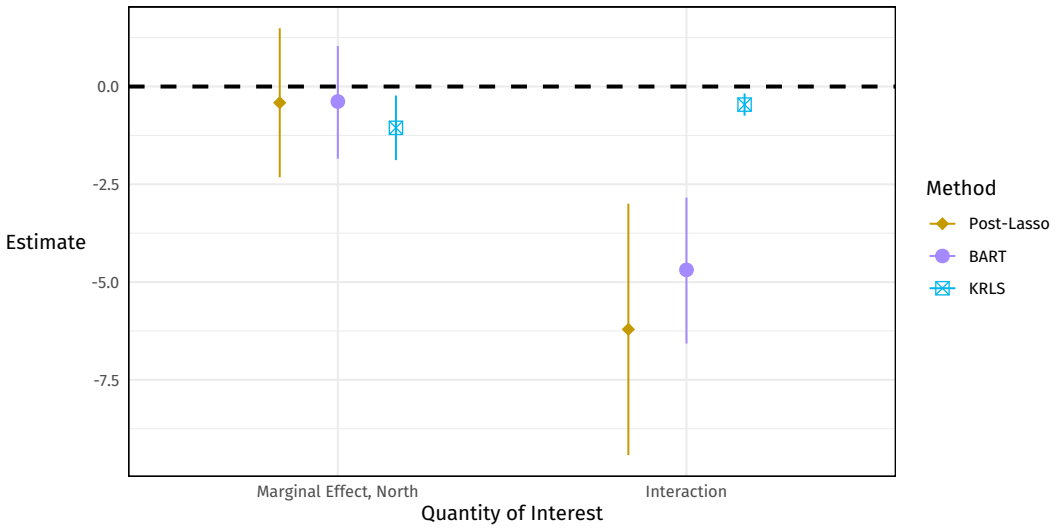

Figure SM.9: Effect of the Direct Primary in the American North and South: Additional Estimators

Estimates from the post-lasso, KRLS, and BART estimators described above. 95% confidence/credible intervals are based on state-clustered standard errors (post-lasso), conventional standard errors (KRLS), and the posterior distribution (BART).

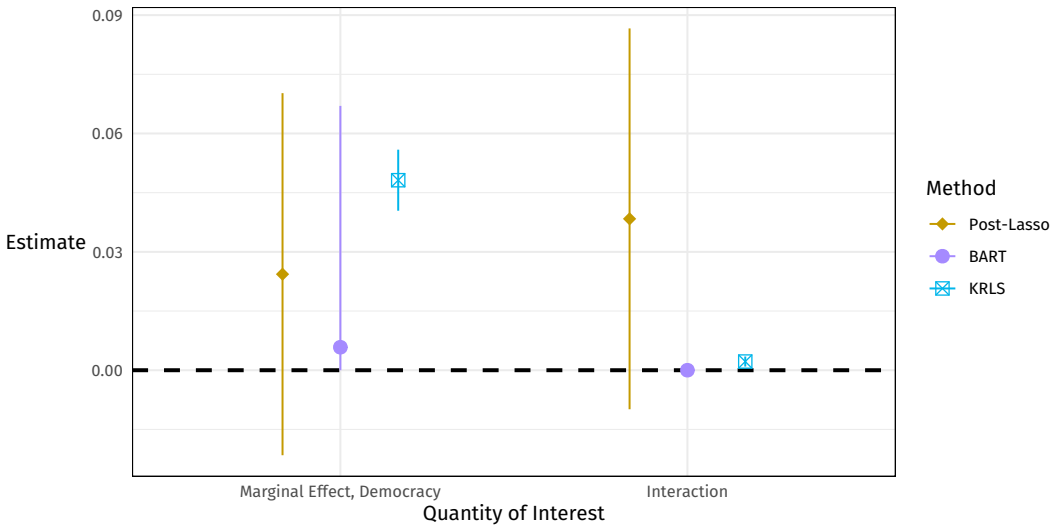

Figure SM.10: Remittances, Protest, and Regime Type: Additional Estimators

Estimates from the post-lasso, KRLS, and BART estimators described above. 95% confidence/credible intervals are based on state-clustered standard errors (post-lasso), conventional standard errors (KRLS), and the posterior distribution (BART).

Bibliography

Belloni, Alexandre, Victor Chernozhukov and Ying Wei. 2016. "Post-Selection Inference for Generalized Linear Models With Many Controls." *Journal of Business & Economic Statistics* 34(4):606–619.