Appendix B: Response Surface C

This online appendix presents simulation results from a third response surface omitted from Section 4 of the article due to space considerations. Further description of the simulation set-up and methods compared is in the primary document. Response surface C is nonlinear and not parallel across treatment groups, with $Y(0) \sim N(Q\beta_{C0}, 1)$ and $Y(1) \sim N(Q\beta_{C1} + \omega_C^s, 1)$, where Q is the matrix of confounding covariates, squared terms for all continuous covariates, and all pairwise interactions. The vector β_{C0} is sampled from (0,1,2) with probabilities (.6,.3,.1) for the coefficients on the original covariates and from (0,.5,1) with probabilities (.8,.15,.05) for the quadratic terms. The vector β_{C1} is sampled identically but independently. The ω_C^s values were chosen in the same way as for Response surface B. The average across simulations of the R^2 values from the linear regression of Y on X is approximately .74.

The results from response aurface C displayed in Figure B.1 are striking. The point estimates from BART are centered on and clustered fairly tightly around the true treatment effect. Estimates from the other methods on the other hand are more unstable with wider variation (RMSE) in estimates. The methods all appear equally capable of identifying that there is a significant effect though BART achieves this with far shorter uncertainty interval lengths on average.

All methods had good coverage when there is complete overlap. Coverage drops for all methods when there is not complete overlap but most dramatically for regression (49%). BART trumps its competitors with regard to estimating heterogeneous treatment effects in both overlap and incomplete overlap settings.

Simulation extensions I also present results from the simulation extensions discussed in Section 5 of the article.

When simulations were run with t-distributed errors the results for all methods worsened a bit in the overlap setting and the differences between the methods were less stark, though BART was still the clear winner. For the incomplete overlap simulations it became difficult to distinguish between BART and the propensity score methods with regard to RMSE though all three still clearly beat linear regression.

As with response surface B, when the standard deviation of the error term was increased to 5 the non-linearities are difficult to detect and at 10 they were virtually impossible to detect. The results of these simulations are available from the author and tell a similar story to the results already presented.

As with response surface A, simulations examining the sensitivity to specification of propensity score model demonstrated basically no impact on the results from response surface C.

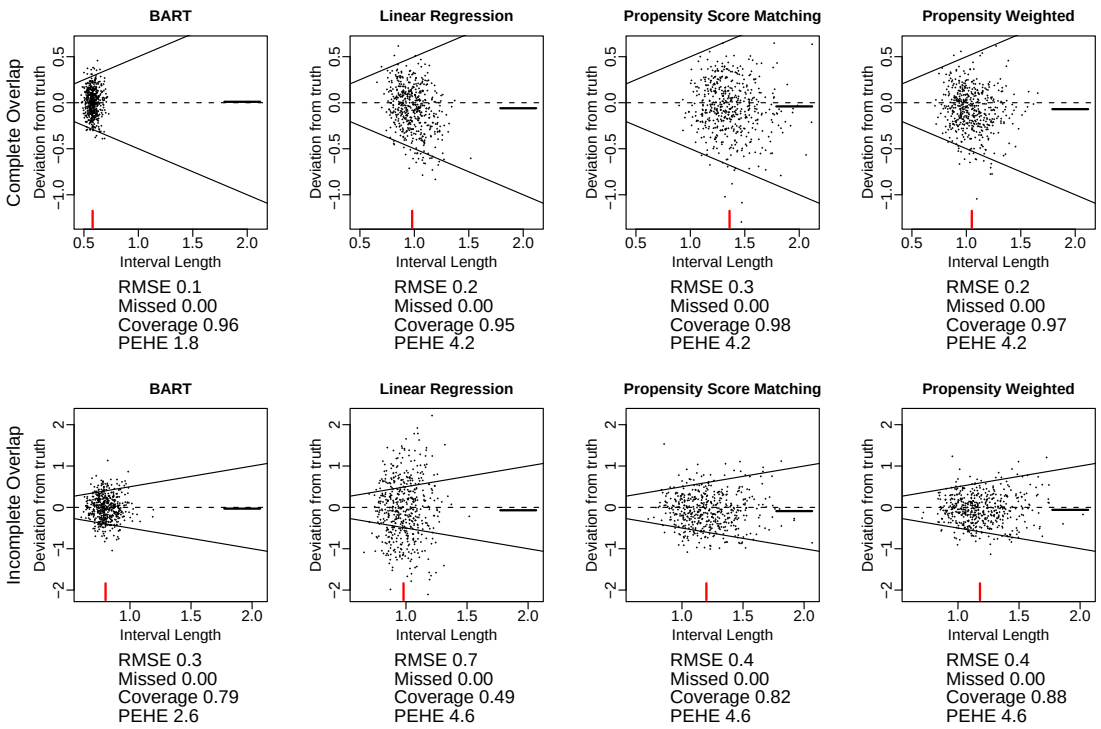

Figure 1: *

Figure B.1: Results from 1000 simulations using nonlinear, not parallel Response surface C. The deviation of the treatment effect estimates from the true effect for 500 simulation draws (randomly sampled from the 1000) are plotted (y-axis) against interval length on the x-axis, with separate plots for each combination of method (columns) and overlap setting: overlap (top row) or incomplete overlap (2nd row). Lines (passing through the origin) with slopes -0.5 and 0.5 display the boundary beyond which a 95% interval corresponding to each point will fail to cover the true parameter value. The short solid line segment on the far right of each plot displays the bias across the 1000 simulations. The short solid line segment at the bottom displays the average interval length. Summary statistics calculated across all 1000 simulations include: root mean squared error (RMSE), % of such intervals that would exclude 0 ("Missed"), and coverage rates for 95% intervals. The final summary statistic, precision in estimation of heterogeneous effects ("PEHE"), reflects the ability to capture individual variation in treatment effects, as described in the text.