BARP!

Notes for MRP reading group

Benjamin Skinner

27 May 2020

Paper

Bisbee, J. (2019). “BARP: Improving Mister P Using Bayesian Additive Regression Trees” *American Political Science Review* 113:4, 1060–1065. doi:10.1017/S0003055419000480

Data

- From five large surveys (following Buttice & Highton, 2013)
    - Three National Annenberg Election Studies
    - Two Cooperative Congressional Election Studies
- 89 items total
- Sample sizes ranging from 25,000 to more than 60,000 observations
- Treat disaggregated state averages as the “ground truth”
- Use covariate strata proportions to poststratify
- Individual-level covariates:
    - sex
    - race
    - age
    - education
- State-level covariates:
    - presidential vote
    - religious conservatism
- Outcomes:

NAES 2000: cutting taxes v. strengthening social security (cbb05), health care spending for uninsured (cbe02), universal health care for children (cbe08), poverty a problem (cbp01), social security spending (cbc01), invest social security in stock market (cbc05), military spending (cbj07), tax rates a problem (cbb01), prescription coverage for seniors (cbe05), right to sue HMOs (cbe14), abortion restrictions (cbf02), death penalty (cbg01), gays in military cbl01) job discrimination (cbl05), school vouchers (cbd02), handgun licenses (cbg05), restrict gun purchases (cbg06), underpunished criminal problem (cbg12), job discrimination (cbm01).

NAES 2004: reduce taxes (ccb13), aid to schools (ccc40), income inequality (ccc41), military spending (ccd03), invest social security in stock market (ccc32), abortion ban (cce01), marriage amendment (cce21), school vouchers (ccc39), gun control (cce31), free trade agreements (ccb82), homeland security spending (ccd57), Patriot Act (ccd67), rebuilding Iraq spending (ccd34), American troops in Iraq (ccd35).

NAES 2008: tax rates-a (cbb01), tax rates-b (cbb01), immigrant path to citizenship (cdd01), border fence with Mexico (cdd04), abortion availability (cea01), same-sex marriage (cec01), environment v. economy (cfb01), American troops in Iraq (cdb01).

CCES 2006: minimum wage (v2072), social security private accounts (v3024), minimum wage (v3072), capital gains tax rates (v3075), taxes v. spending (v4040), taxes v. spending v. borrowing (v4044), abortion (v3019), late-term abortion (v3060), stem cell funding (v3063), illegal immigrant citizenship (v3069), environment (v3022), affirmative action (v3027), free trade—CAFTA (v3078), military use—oil supply (v3029), military use—terrorist camps (v3030), military use—genocide (v3031), military use—spread democracy (v3032), military use—protect allies (v3033), military use—help UN (v3034), Iraq troop withdrawal (v3066).

CCES 2008: balanced budget (cc309), privatizing social security (cc312), minimum wage (cc316b), health insurance for children (cc316e), assistance for housing crisis (cc316g), taxes v. spending (cc420), abortion (cc310), stem cell research (cc316c), gay marriage (cc316f), jobs v. environment (cc311), affirmative action (cc313), eavesdropping without court order (cc316d), free trade-NAFTA (cc316h), bank bailout (cc316i), carbon tax (cc422), Iraq troop withdrawal (cc316a), military use—oil supply (cc418_1), military use—terrorist camps (cc418_2), military use—genocide (cc418_3), military use—spread democracy (cc418_4), military use—protect allies (cc418_5), military use—help UN (cc418_6), internet absentee voting (cc419_1), election day registration (cc419_2), voter eligibility (cc419_3), vote by mail (cc419_4), automatic registration (cc419_5), photo ID to vote (cc419_6).

From Buttice and Highton, 2013, p. 465

Method

- Randomly sample with replacement sample sizes:
  - small (1,500)
  - medium (3,000)
  - large (4,500)
- 200 replications
- To fit models:
  - MRP: `lme4`
  - BARP: custom BARP package (based on `bartMachine` package)
  - Other (see below): `SuperLearner`

- Other comparison (from Section 7 of supporting information):
    - Gradient Boosting Machines (GBM)
    - Kernel k-Nearest Neighbor (kNN)
    - Neural Networks (NNET)
    - Support Vector Machines (SVM)
    - Elastic-net regularization (LASSO)
- Evaluation metrics
    - mean absolute error
    - correlation between predicted state values and "true" values
    - Buttice & Highton (2013) use standardized versions of these values; Bisbee shows this version in Figure 1 of supporting information — results are more strongly in favor of BARP over MRP

Results

- Predictive accuracy (figure 1):
    - MAE lower with BARP
    - ICC higher with BARP
- Sensitivity to misspecification (figure 2):
    - Drop state-level covariates from model and perform t-test on difference
    - BARP appears less sensitive to misspecification
- Sensitivity to within-state sample size (figure 3):
    - BARP less sensitive
- Comparison with other regularization methods (takeaway from supporting information, p. 18):

> These comparisons produce three general conclusions.
>
> - First, MRP is competitive with more sophisticated methods in terms of mean absolute error in larger sample sizes.
> - Second, BARP is among the best-in-class in terms of mean absolute error but is more in the middle of the pack in terms of interstate correlation.
> - Third, the best methods evaluated in terms of mean absolute error are not the same as the best methods judged by interstate correlation. Across both metrics, LASSO regularization is competitive.

Takeaway

- When conditions are good (strong 2nd-level covariates and more variation between groups than within), MRP and BARP are about the same.
- When less properly specified, BARP may be less sensitive to problems than MRP

An aside to Buttice and Highton (2013)

Buttice, M., & Highton, B. (2013). How Does Multilevel Regression and Poststratification Perform with Conventional National Surveys? Political Analysis, 21(4), 449-467. www.jstor.org/stable/24572674

> Given what researchers have learned about measurement error and reliability, few contemporary public opinion scholars would rely on national samples of 1500 to estimate state or congressional district opinion by simply computing the DM estimates (Erikson 2006). As a consequence, our focus is on how well MRP performs in an absolute sense. In the case of national surveys of typical size, because DM performs so poorly, the fact that MRP outperforms DM does not imply that MRP performs *well* and that its estimates should be employed in substantive analyses. (Buttice & Highton, p. 453)

They expect MRP performance (public opinion across states) to be a function of (p. 453–454):

1. Degree to which individual-level covariates account for opinion
2. Degree to which state-level covariates account for opinion
3. Degree to which true opinion varies across states relative to within states

Table 2 Monte Carlo results

	Condition			
Strength of state covariates	*Low*	*Low*	*High*	*High*
Population ICC	*Low*	*High*	*Low*	*High*
Correlation with true values				
Average	0.17	0.61	0.66	0.87
Percent of samples < 0.50	99	8	17	0
Percent of samples 0.50 – 0.75	3	90	36	0
Percent of samples > 0.75	0	2	47	100
Absolute standardized bias				
Average	0.71	0.45	0.38	0.34
Percent of samples < 0.50	46	64	70	70
Percent of samples 0.50 – 0.75	12	18	16	24
Percent of samples > 0.75	42	18	14	6
Absolute standardized error				
Average	0.92	0.62	0.67	0.42
Percent of samples < 0.50	0	1	10	97
Percent of samples 0.50 – 0.75	3	98	63	3
Percent of samples > 0.75	97	1	27	0
90% confidence interval coverage				
Average	0.85	0.80	0.91	0.79
Percent of states < 0.85	40	32	20	42
Percent of states 0.85 – 0.95	38	16	48	28
Percent of states > 0.95	22	52	32	30

Note. For each of the four combinations of strength of state covariates and population ICC, five hundred trials were conducted based on sample sizes of 1500.

*From Buttice & Highton, 2013, p. 463. (**NOTE:** Meaning of ICC is different from ICC in Bisbee)*

> Across the three measures, then, the Monte Carlo results show that MRP performance is neither uniformly strong nor uniformly weak. The quality of the estimates depends crucially on how well the state-level covariates account for opinion variation across the states along with how the interstate differences in opinion compare to the intrastate differences. (Buttice & Highton, p. 462)

General Questions

- How much credence should we give to “horse race”-style comparisons using real data, particularly when the ground truth is just assumed from the data?

> Initially, it may seem problematic to treat a sample as the population, but several factors suggest otherwise. First, consider an item on abortion opinion for which there are ninety-five Vermonters in a national sample of 30,000 respondents. We treat those ninety-five as the Vermont population. If they happen to be wildly unrepresentative, then some state-level covariates included in the multilevel opinion model (like state presidential vote) will not perform well, but this will be reflected in the measure of state-level covariate strength. To the extent that state-level covariate strength matters (and below we show that it is very important), the problem will be addressed when we analyze the relationship between the strength of the state-level covariates and MRP performance. Second, in supplemental analyses one of the state-level covariates we included in the multilevel opinion model was state ideology, which was measured as the disaggregated state mean ideology for the "population." If the ninety-five Vermonters are not typical Vermonters and exhibit abortion opinion that is not truly representative of Vermonters, then the same will likely be true for their ideological preferences. So, by including "true" ideological preferences as a state-level covariate in the MRP models, we account for this. As it turns out, for the eighty-nine items, overall MRP performance and the correlates of MRP performance were nearly identical when we included state ideology and when we did not. (Buttice & Highton, p. 455, footnote 14)

- Differences due to use of `lmer()` versus full Bayes via Stan?