Validating Ecological Inference Estimates from Cast Vote Records

Shiro Kuriwaki (Harvard University)

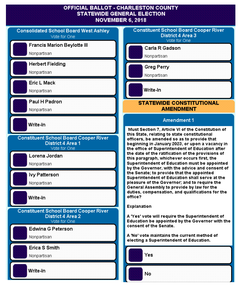
July 2019

https://www.shirokuriwaki.com

Data: Cast Vote Records (CVRs)

CVRs are electronic copies of voted ballots, and measure ticket splitting rates exactly

☑ Individual-level
☑ Actual vote choice observed without error
☑ Down-ballot races observed



South Carolina's ES&S iVotronic (DRE)

```
RUN DATE:11/09/18 02:32 PM ELECTION ID: 10110618
PRECINCT 537 - St Andrews 37
*    Henry McMaster          Governor
     Mark Hammond            Secretary of State
     Rosalyn Glenn           State Treasurer
     Alan Wilson             Attorney General
     Richard Eckstrom        Comptroller General
     Molly Mitchell Spearman State Superintendent of Education
     Hugh Weathers           Commissioner of Agriculture
     Joe Cunningham          CON0001 U.S. House of Representativ
     Lin Bennett             HOU0114 State House of Representati
     Irv Condon              Probate Judge
     Tom Hartnett            Register of Deeds
     Hal Hanvey              Soil and Water District Commission
     Kate Darby              CSB East Cooper
     Jake Rambo              CSB East Cooper
     Vivian Sheppard Pettigrew CSB North Area
     Paul H Padron           CSB West Ashley
     Cullen Baney            SCH0010 Con SB St Andrews District
     Daron Lee Calhoun II    SCH0010 Con SB St Andrews District
     Christopher Mahon       SCH0010 Con SB St Andrews District
     No                      Statewide Constitutional Amendments
*    Republican              STRAIGHT PARTY
     Henry McMaster          Governor
     Mark Hammond            Secretary of State
     Curtis Loftis           State Treasurer
     Alan Wilson             Attorney General
```

South Carolina Data from Kuriwaki (2019), "Party Loyalty on the Long Ballot: Is Ticket Splitting More Prevalent in State and Local Elections?" (https://doi.org/10.31235/osf.io/bvgz3)

Estimands

Goal: Using voteshares from $i \in 1, ..., N_p$ precincts, infer the unobserved **Column Proportions**.

	Presidential Vote		
Down-ballot Vote	Republican (Trump / Romney)	Democrat (Clinton / Obama)	(Voteshare)
Republican	W_{straight}	W_{split}	Y_R
Democrat	$1 - W_{\text{straight}}$	$1 - W_{\text{split}}$	$1 - Y_R$
(Voteshare)	X_R	$1 - X_R$	1 (= n_i voters)

Final Estimand: total proportion of

straight ticket voting among Presidential Republicans and split ticket voting among Presidential Democrats

$$\sum_{i=1}^{N_p} n_i X_{R,i} W_{\text{straight},i} \Big/ \sum_{i=1}^{N_p} n_i X_{R,i} \quad \text{and} \quad \sum_{i=1}^{N_p} n_i (1 - X_{R,i}) W_{\text{split},i} \Big/ \sum_{i=1}^{N_p} n_i (1 - X_{R,i})$$

Methods Compared

Homogeneous Precincts

- Use Y_R from precincts where $X_R \approx 0$ or $X_R \approx 1$ only

Ecological Regression

- Extrapolate from correlation of aggregate marginals
- Regress Y_R on X_R and $1 - X_R$

Wakefield 2 × 2

- Model counts n_{X_R,Y_R}
- $n_{X_R,Y_R} \sim \text{Binom}(n_{X_R}, \mu)$,
- $\mu \sim$ Normal

Imai, Lu, Strauss 2 × 2

- Model proportions W
- $\text{logit}^{-1}(\boldsymbol{W}) \sim$ Normal
- $\mu \sim$ Normal; $\Sigma \sim$ InvWish

Rosen, King, Jiang, Tanner 4 × 4

- Model proportions W
- Model 4 options: Republican, Democrat, Other, Abstain
- $(W_1, W_2, W_3, W_4) \sim$ Dirichlet
- Prior: $\boldsymbol{X} \sim \text{Multi}(\theta_1, \theta_2, \theta_3, \theta_4)$

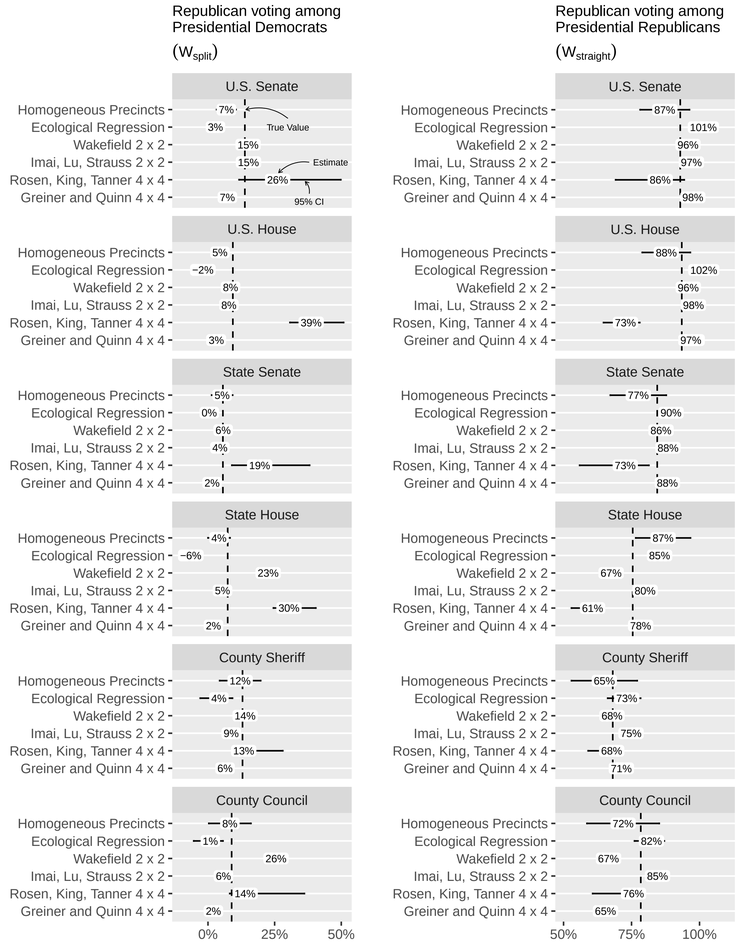
Greiner and Quinn 4 × 4

- Model counts n_{X_R,Y_R}
- $(n_{X_1}, n_{X_2}, n_{X_3}, n_{X_4}) \sim$ Multi
- Also model 4 × 4
- $\mu \sim$ Normal; $\Sigma \sim$ InvWish

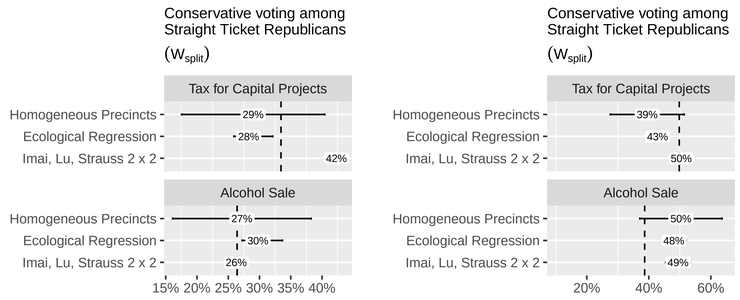
Do Ecological Inference Methods Recover the Ground Truth?

Most methods over-estimate straight ticket rates, and homogeneous precincts does about as well as EI methods.

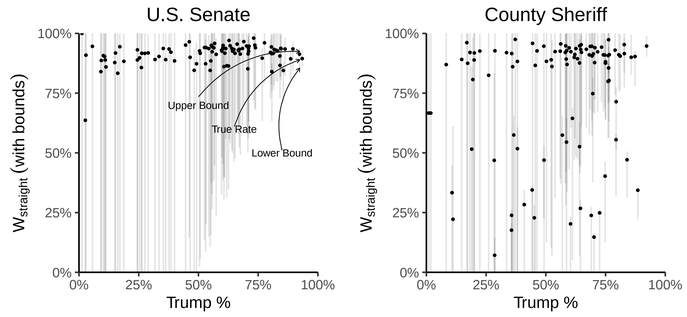
Voting on Candidates



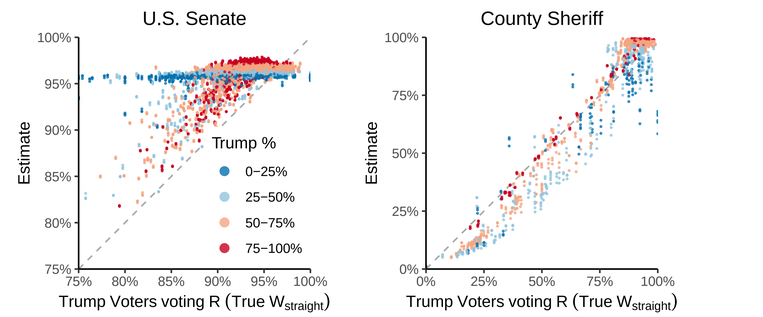
Voting on Referendums



Deterministic Bounds are Uninformative



Small Precincts Drive Error Rates



Possible Adjustments: Race as a Contextual Variable