Smoothing Mixed Effects Models and Election Poll Aggregation: An Application to the 2020 Democratic Primary

Lisa Wilson

June 8, 2020

Background

Motivation

- Election polling and forecasting models under more scrutiny after the 2016 U.S. election
- Different election forecasting approaches
    - Fundamental models: historical data, economic indicators, other covariates; no current polling data
    - Poll aggregation models: combines current polling data, sometimes historical as well
    - Hybrid/synthetic models: incorporate fundamental elements and poll aggregation

Wright's Smoothing Mixed Effects Model

- Wright (2018) evaluates models used by FiveThirtyEight, NYT Upshot, Princeton Election Consortium, and HuffPost
- Proposes a smoothing mixed effects model with the log-ratio of percent support for Clinton over Trump in each state as the response, using only state polls
  - Matches or outperforms the prediction sites, predicts a higher probability of a Trump electoral college win
- Log-ratio is less affected by the percentage of third-party supporters, which tends to be higher in polls than in election results
- Log-ratio and the difference between candidate support have the same sign, so predictions about which candidate will win a particular state can be made on either scale

Model Details

$$y_{ij}(t) = \mu(t) + v_i(t) + \epsilon_{ij}$$

- $y_{ij}(t) = \log(\frac{A_{ij}(t)}{B_{ij}(t)})$: log-ratio of percent support for candidate A to candidate B for state i and poll j at time t days before the election
- $\mu(t)$: national fixed effect (as measured from state polls)
- $v_i(t)$: state-level random effect
- ϵ_{ij}: iid normal error terms with mean 0

Model Details

$$y_{ij}(t) = \mu(t) + v_i(t) + \epsilon_{ij}$$

- $\theta_i(t) = E(y_i(t))$: true log-ratio of support for candidate A to candidate B
- $\hat{\theta}_i(0)$: spline-extrapolated estimate of the log-ratio of candidate support on election day in state i
- t: number of days between the mid-date of the poll and the election

Model Details

- Model estimation with the R package `sme`
  - Fits smooth functions $\mu(t)$ and $\{v_i(t)\}$ with splines using maximum likelihood estimation, penalized by bandwidth parameters λ_μ and λ_v
- Use a training and test set of polls to determine the λ_μ, λ_v combination that minimizes RMSE for the test set
  - All possible combinations of $\lambda_\mu, \lambda_v = (0.5, 1, 5, 10, 50, 100, 500, 1000, 5000)$
- To obtain $\hat{\theta}_i(0)$, use spline extrapolation with model estimates for $\mu(t)$ and $\{v_i(t)\}$ at all time points present in data

Model Details

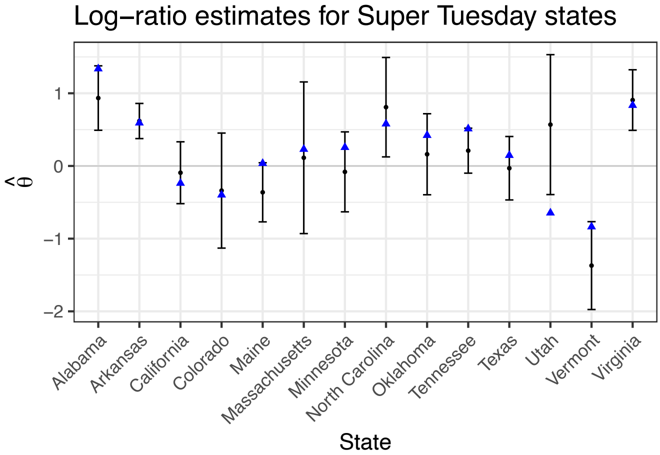
- Calculate standard errors by using a parametric bootstrap method to estimate the variance-covariance matrix for $\hat{\theta}_i(0)$
  - Generate new values of log-ratio of percent candidate support for state i and poll j for each pollster in each state
  - Obtain $\hat{\theta}_i(0)$
  - Repeat for 500 sets of estimates
  - Take the covariance of the matrix of $\hat{\theta}_i(0)$ to get the estimated variance-covariance matrix

Application

Super Tuesday 2020 Data

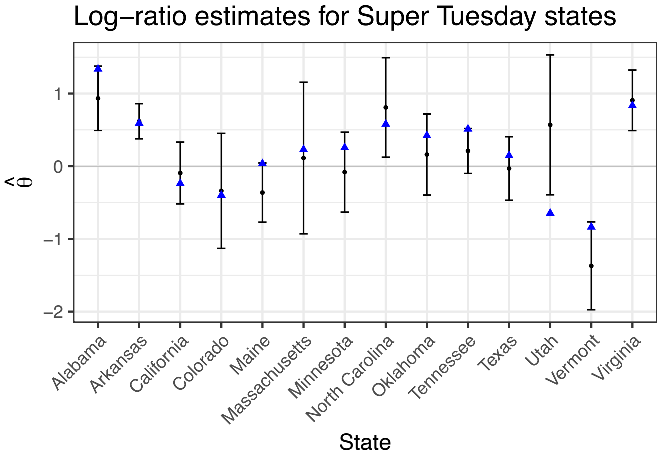
- Focusing on the 2020 Democratic presidential primary
- Super Tuesday: March 3, 2020
  - Alabama, Arkansas, California, Colorado, Maine, Massachusetts, Minnesota, North Carolina, Oklahoma, Tennessee, Texas, Utah, Vermont, and Virginia
- Biden and Sanders chosen as two candidates to model
  - Frontrunners of the moderate and progressive wings
- Raw polling data from FiveThirtyEight's poll tracking site
- $\lambda_\mu = 5000$, $\lambda_v = 100$ identified as best bandwidth values

Results



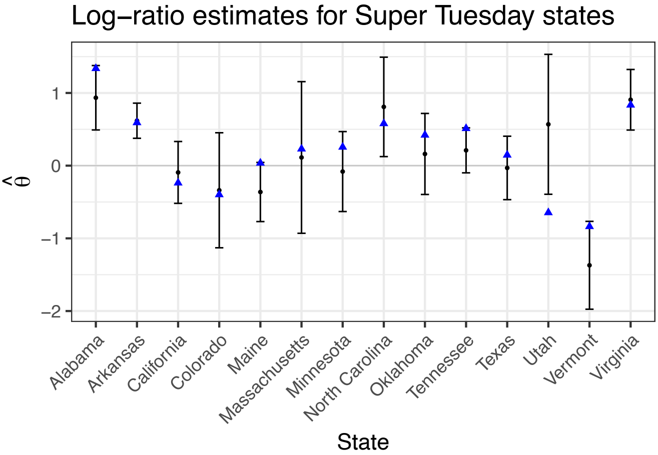
- Positive: more support for Biden; Negative: more support for Sanders

Results



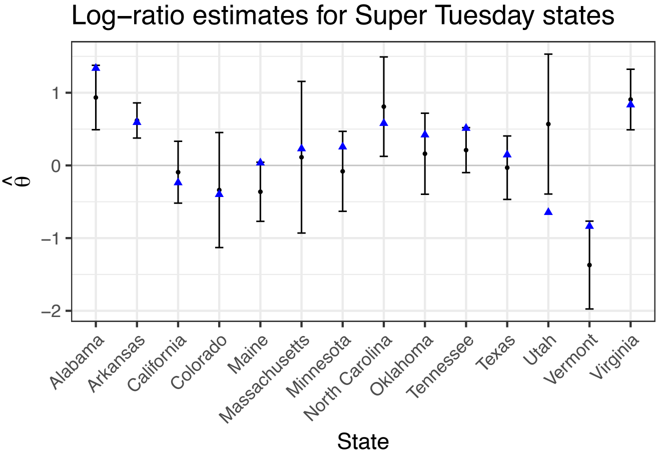
- Maine, Minnesota, Texas, and Utah were miscalled

Results



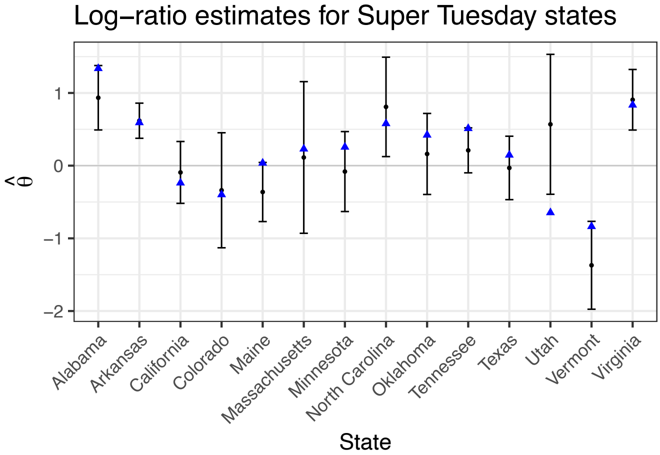
- Arkansas, Colorado, and Virginia closest; Utah furthest

Results



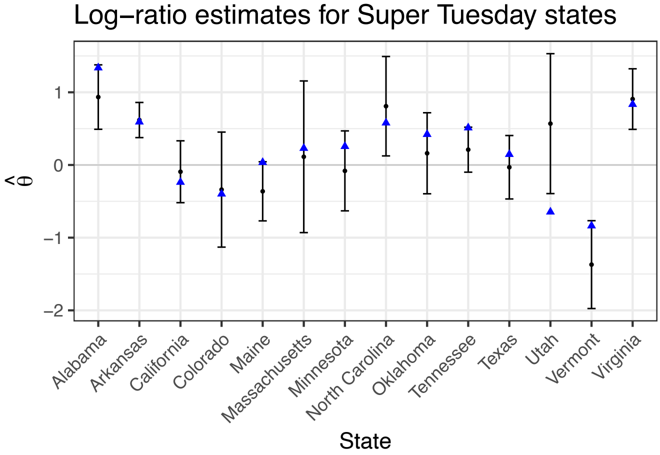
- RMSE $= 0.4210$

Results



- Coverage rate $= 92.86\%$

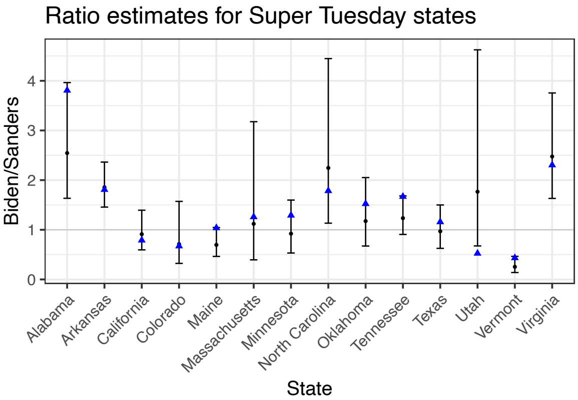
Results



- Intervals for 9 out of 14 states included 0, indicating a Biden or Sanders win was plausible

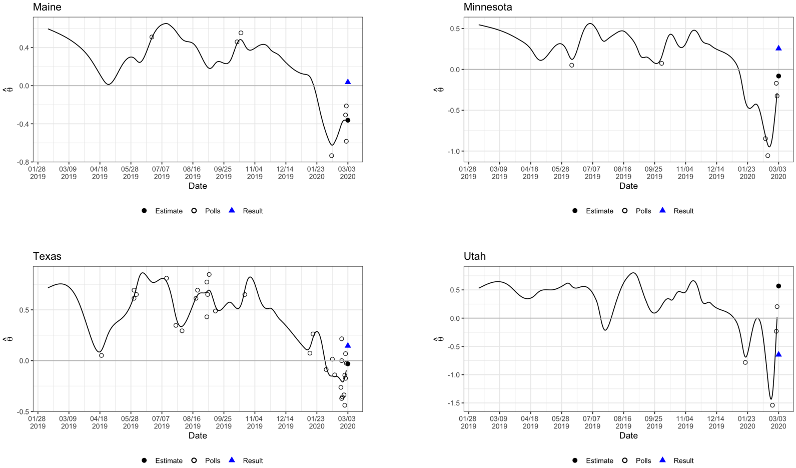
Results

- Not appropriate to transform log-ratio to difference scale, but can transform to ratio scale



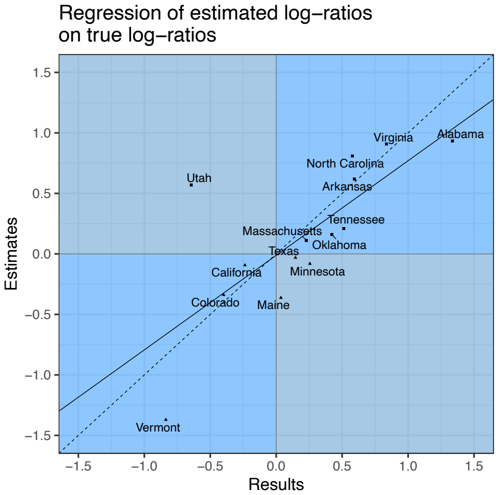
- RMSE $= 0.5375$

Results

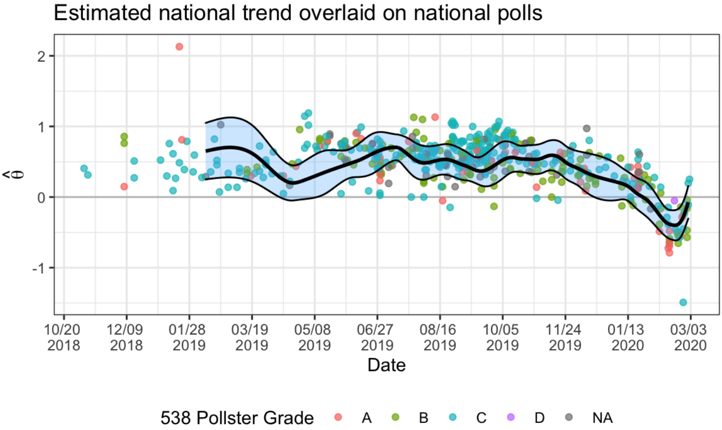

Maine
Minnesota
Texas
Utah
$\hat{\theta}$
Date
01/28 2019
03/09 2019
04/18 2019
05/28 2019
07/07 2019
08/16 2019
09/25 2019
11/04 2019
12/14 2019
01/23 2020
03/03 2020
0.4
0.0
-0.4
-0.8
0.5
-0.5
-1.0
-1.5
Estimate
Polls
Result

Results

$$\hat{\theta}_i(0) = -0.012 + 0.782\theta_i(0)$$



Results



- RMSE $= 0.2922$

Results

Comparison to FiveThirtyEight predictions

	SME	FiveThirtyEight
RMSE	0.421	0.3349
Correlation	0.7441	0.8688
$\hat{\beta}_0$	-0.0117	-0.138
$\hat{\beta}_1$	0.7816	0.7885
Miscalled	ME, MN, TX, UT	ME, MA, MN

Conclusion

Conclusions

- Smoothing mixed effects model works well in this primary setting, in addition to general election setting demonstrated by Wright
- Limitations
  - Model may have performed well because primary had become essentially a two-candidate race
    - Possible Massachusetts counterexample: votes more evenly split among Biden, Sanders, and Warren
  - Also possible that model performs well for an election day with many states voting, but may not work as well with fewer states
  - Can only interpret on the log-ratio or ratio scale

Conclusions

- Strengths
  - Simplicity: only state polls available before election day
  - Still indicates relative support for each frontrunner in each state, predicts winner
  - Works even for states with few polls because it can draw on other states, estimated national effect
- Future work
  - Extending the model to analyze relative support among three or more candidates

Thank you!

Questions?