Using Poisson Binomial Models to Reveal Voter Preferences

Evan Rosenman

Stanford University, Department of Statistics

Overview

Secret ballot obfuscates individual voters' choices. But publicly reported data include:

- Covariates for every voter (*voter file*)

Voter Name	County	Precinct	Gender	Age	…

- Precinct-level vote tallies for every candidate (*election results*)

County	Precinct	# of votes for Dem	# of votes for Rep

Ecological inference problem: seek to model individuals, but only have aggregate outcomes.

Contributions:

- Propose model structure and develop approximate algorithm finding MLE
- Demonstrate comparative efficacy on problem of revealing voter preferences

Ecological Inference

Problem has a rich history in the literature:

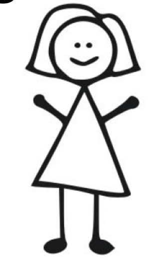
- In **poli sci**, early work related to the VRA. Later work: Wakefield (2004), Jackson (2008)
- Recent interest from **machine learning**: Flaxman (2016), Patrini (2014), Rueping (2010)

Problem is increasingly relevant in other settings, such as **differential privacy.**

Thanks to the NDSEG Fellowship and to Google for their support. Thanks also to Nitin Viswanathan, Pete Mohanty, Art Owen, & Michael Baiocchi for collaboration and feedback.

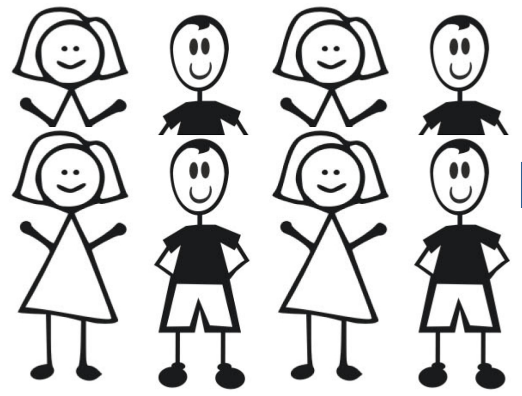
Poisson Binomial Model

Voters j in precinct i are modeled as independent but not identically distributed *Bernoulli(p_{ij})* variables, with **logistic regression** formulation for modeling p_{ij}:



$$p_{ij} = \sigma(\boldsymbol{x}_{ij}^{\mathsf{T}}\beta)$$

Precinct-level Democratic votes D_i are modeled as independent **Poisson Binomial** variables (sum of independent Bernoullis):



$$\ell_i(\beta) = \sum_{A \in F_{D_i}} \prod_{j \in A} p_{ij} \prod_{j \in A^c} (1 - p_{ij})$$

F_{D_i} = set of configurations of D_i Democratic votes for precinct i

A = set of voters voting Democrat in this configuration

A^c = set of voters voting Republican in this configuration

Finding Coefficients

Want to obtain **MLE** for β, but

- Can't efficiently compute Poisson binomial probabilities
- True likelihood is non-convex

Approach

- Approx. likelihood via Lyapunov CLT

$$D_i \xrightarrow{d} N\left(\sum_{j \in S_i} p_{ij}, \sum_{j \in S_i} p_{ij}(1 - p_{ij})\right)$$

$$\ell_i(\beta) \approx -\log(\phi_i) + \frac{1}{\phi_i^2}(D_i - \mu_i)^2$$

- Train model via batch gradient descent

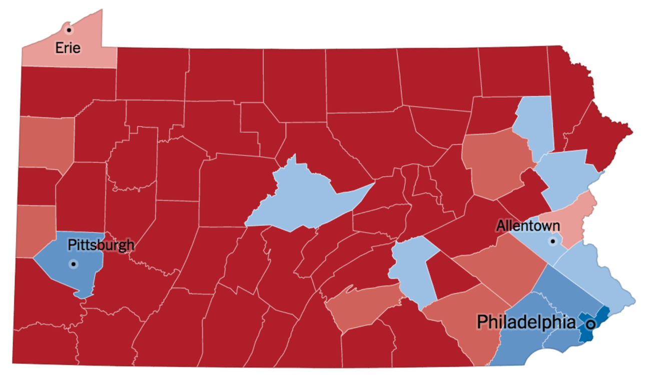
$$\nabla_\beta \ell_i \approx \frac{1}{\phi^2}(D_i - \mu_i)\left(\sum_j p_{ij}(1 - p_{ij})\boldsymbol{x}_{ij}\right) -$$

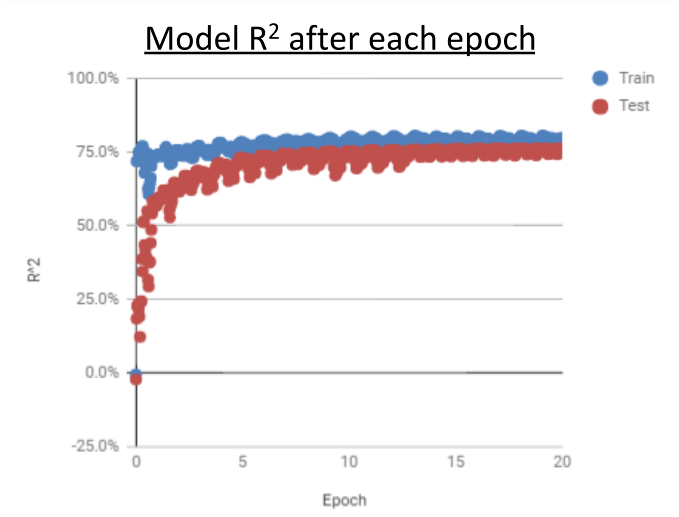
$$\frac{1}{2}\left(\frac{(D_i - \mu_i)^2}{\phi_i^4} - \frac{1}{\phi_i^2}\right)\left(\sum_i (2p_{ij} - 1)(1 - p_{ij})p_{ij}\boldsymbol{x}_{ij}\right)$$

Analysis: PA 2016

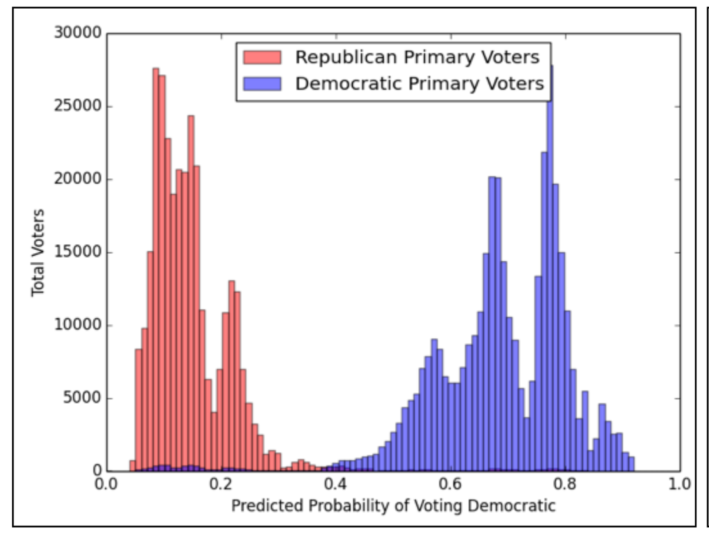
Deployed methods to model results of 2016 presidential election in Pennsylvania.

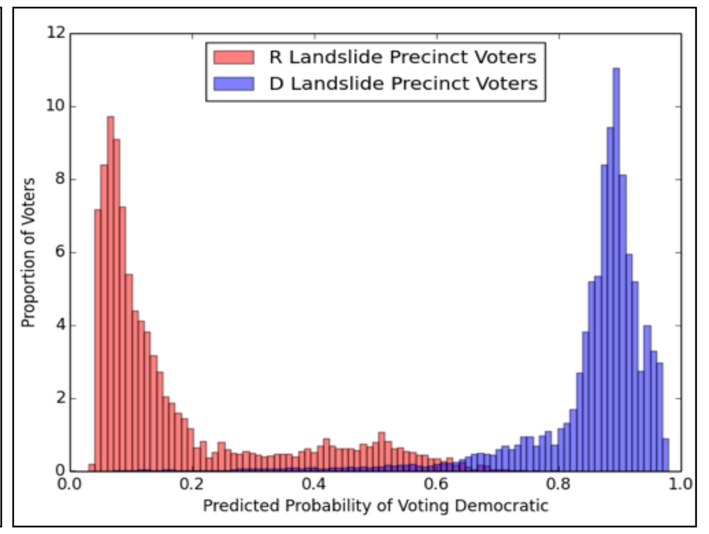
Able to merge data comprising 71% of voters (4.37MM)





Because we lack labeled data, we validate using weak labels. Model produces expected bimodal distribution in both cases.





Model coefficients after training on entire dataset

(Positive = more likely to have voted for Clinton)

Female	*0.13*	Dem Primary Voter	*0.33*
Male	*−0.21*	Rep Primary Voter	*−0.42*
Age	*−0.53*	County White %	*−0.14*
Apartment Dweller	*1.29*	County Black %	*0.09*
Registered Dem	*1.00*	County College %	*0.28*
Registered Rep	*−1.23*	County Income	*−0.03*

Performance Bake-Off

To analyze performance vs. competitor techniques, we use a related task – predicting whether someone votes – for which we have individual-level outcomes.

We source data from Morris County, NJ and aggregate voter tallies to the precinct level. We compare ROC AUC values on a holdout set for our techniques vs. competitor ecological inference methods.

	Demographics and Voting History			
	2017	**2016**	**2015**	**2014**
Standard Methods (non-ecological)				
Logistic Regression	85.9%	84.5%	88.6%	89.5%
GBM	86.2%	85.5%	88.8%	89.6%
Proposed Methods				
Logit with Gaussian Gradient	<u>83.9%</u>	<u>82.0%</u>	<u>81.0%</u>	86.3%
Logit with Gaussian Gradient, PoiBin Backtracking	83.8%	82.0%	81.0%	<u>86.4%</u>
Logit with Gaussian Gradient, PoiBin Backtracking, True Gradient	83.8%	81.9%	80.6%	86.3%
Neural Net with Gaussian Gradient	72.1%	76.8%	80.4%	74.1%
Comparison Methods				
Logistic Regression on Aggregates	75.0%	72.4%	77.2%	76.8%
Ecological Regression	67.5%	68.7%	71.8%	76.1%
Inverse Calibration	64.2%	77.6%	78.4%	66.9%
Mean Map	45.4%	54.4%	48.4%	51.8%
Laplacian Mean Map	49.5%	51.5%	57.6%	49.4%
Alternating Mean Map	51.9%	52.9%	44.4%	46.2%

Using with a rich covariate set, our methods outperform competitor ecological inference methods, and nearly match methods with access to individual-level outcomes.

Future Directions

Current work can be found on **arXiv:**

- "Using Poisson Binomial GLMs to Reveal Voter Preferences" (*1802.01053 – with Nitin Viswanathan*)
- "Some New Results for Poisson Binomial Models" (*1907.09053*)

Future directions of research:

- Define conditions for existence of a **finite MLE**
- Developing **valid confidence intervals** for coefficients
- Extend to more **flexible models** for probabilities p_{ij}