Prediction of the 2018 Midterm Elections

Jiafeng Chen* Joon Yang†

November 3, 2018

Let

$$Y_i = \frac{\text{Republican\%}}{\text{Republican\%} + \text{Democrat\%}}$$

be the proportion of Republican vote share between the major parties in district i.[1] Let x_i be a list of features for the district. Assume the (misspecified) linear probability model $Y_i \sim \mathcal{N}(\mu_{i0}, \sigma_0^2)$, where $\mu_{i0} = x_i^T \beta_0$. The model is misspecified since $Y_i \in [0,1]$ but Normal is supported on $\mathbb{R}$. We justify the misspecification by noting that $\text{Beta}(a,b) \xrightarrow{d} \mathcal{N}(\mu, \sigma^2)$ where $a, b \to \infty$ in such a way that expectation and variance are fixed at μ, σ^2.[2] Let

$$\widehat{\beta}_0 = \arg\min_{\beta} \sum_{i=1}^{N} (y_i - x_i^T \beta)^2 + \lambda \left(\alpha \left\| \beta \right\|_1 + (1-\alpha) \left\| \beta \right\|_2 \right)$$

be fitted with an elastic net regularizer over the training data, where λ is chosen via K-fold cross validation and α is some fixed constant, say 0.9. Let $\widehat{\sigma}_{0i}^2 = \sum (y_j - x_j^T \widehat{\beta}_0)^2$ be fitted as the variance of the residuals, where the sum could be over districts in the same state, over all districts, or some kernel-weighted estimator. For simplicity, we stratify variance estimate by state.[3] Let $\widehat{\mu}_{i0} = x_i^T \widehat{\beta}_0$.

For a district that corresponds to an upcoming election, we form a prior $Y_i \sim \mathcal{N}(\widehat{\mu}_{i0}, \widehat{\sigma}_0^2)$, or in vector form

$$Y \sim \mathcal{N}(\widehat{\mu}_0, \widehat{\Sigma}_0)$$

Note that in such a formulation, we ignore the sampling variance of $\widehat{\mu}_{i0}$ and $\widehat{\sigma}_0^2$,[4] instead forming a plug-in estimate, appealing to the law of large numbers.[5] To obtain a more accurate and timely prediction for the district, we update the prior in two steps.

First, to take into account the "blue wave," we update our prior via the generic congressional ballot.[6] Formally, we model generic congressional poll as $Z_G \mid Y_i \sim \mathcal{N}(Y_i, \sigma_G^2)$, where σ_G^2 is

*Harvard College, jiafengchen@college.harvard.edu

†Harvard College, joonhyukyang@college.harvard.edu

[1] Assuming districts have no time dimension: i.e. Alabama-01 is represented by different i, j's across two different years. We also assume for simplicity that third parties never win elections, which seems accurate in the case for 2018.

[2] Alternatively, we could fit some generalized linear model with link function implied by a Beta distribution, but Normal linear model provides a lot of computational ease.

[3] One could also estimate variance over a holdout set, which might improve bias.

[4] The elastic net regularizer in the fitting method for β makes the sampling variance of $\widehat{\beta}_0$ difficult to compute.

[5] From this point on, we drop the hat on μ_{i0}, σ_0.

[6] https://projects.fivethirtyeight.com/congress-generic-ballot-polls/

estimated from the 90% confidence interval provided by FiveThirtyEight. We have the following data generating process:

$$\begin{aligned} Y &\sim \mathcal{N}(\widehat{\mu}_0, \widehat{\Sigma}_0) \\ Z_G \mid Y &\sim \mathcal{N}\left(n^{-1} 1^T Y, \sigma_G^2\right) \end{aligned}$$

We update our prior to form an intermediate posterior:

$$Y \mid Z_G \sim \mathcal{N}(\mu_1, \Sigma_1).$$

From now on, we drop the conditioning on Z_G.

Second, for some districts, we observe a number of district-specific polls $Z_{i1}, \ldots, Z_{iJ_i}$.[7] The variance across polls is much higher than implied by a simple Beta-Binomial model, where one assumes that each poll is an independent $\text{Bin}(n_j, p)$ where p is sampled from a Beta (again, approximately Normal) prior. As such, we hesitate from using a simple Beta-Binomial updating procedure and opt for the following model: We assume that poll j has an independent bias $\epsilon_j \sim \mathcal{N}(0, \sigma_p^2)$, where σ_p is estimated as the empirical variance of poll outcomes in a district:

$$\begin{aligned} Y \sim \mathcal{N}(\mu_1, \Sigma_1) \quad \epsilon_j \sim \mathcal{N}(0, \sigma_p^2) \\ Z_{ij} \mid Y_i, \epsilon_j \sim n^{-1} \cdot \text{Bin}(n, e_i^T Y + \epsilon_j), \end{aligned}$$

where e_i is the ith standard basis vector. It's somewhat difficult to compute the posterior $Y \mid Z_{i1}, \ldots, Z_{iJ}$. Instead we may assume the misspecified model, justified as $n^{-1}\,\text{Bin}(n, p)$ is approximately Normal when n is large.

$$\begin{aligned} Y &\sim \mathcal{N}(\mu_1, \Sigma_1) \\ Z_{ij} \mid Y &\sim \mathcal{N}\left(e_i^T Y, \frac{1}{4n_j} + \sigma_p^2\right), \end{aligned}$$

so as to (a) take advantage of Normal-Normal conjugacy and (b) ignore the dependence of $\mathbb{V}(Z_{ij})$ on Y_i. We now have the posterior by, say, sequentially Bayesian updating:

$$Y \mid (Z_{i1}, \ldots, Z_{iJ_i})_{i=1}^n \sim \mathcal{N}\left(\mu_2, \Sigma_2\right).$$

We predict $\widehat{Y}_i = \mu_{2i}$, and, naturally

$$\widehat{\text{Winner}}_i = \begin{cases} \text{Republican} & \mu_{2i} > .5 \\ \text{Democrat} & \mu_{2i} < .5 \end{cases}$$

[7] In practice, we take the J most recent polls (if available) in district i, where $J = 10$.