Bayesian Combination of State Polls and Election Forecasts*

Kari Lock[1] and Andrew Gelman[2]
[1] *Department of Statistics, Harvard University, lock@stat.harvard.edu*
[2] *Department of Statistics and Department of Political Science, Columbia University, gelman@stat.columbia.edu*

25 November 2008

Abstract

A wide range of potentially useful data are available for election forecasting: the results of previous elections, a multitude of pre-election polls, and predictors such as measures of national and statewide economic performance. How accurate are different forecasts? We estimate predictive uncertainty via analysis of data collected from past elections (actual outcomes, pre-election polls, and model estimates). With these estimated uncertainties, we use Bayesian inference to integrate the various sources of data to form posterior distributions for the state and national two-party Democratic vote shares for the 2008 election. Our key idea is to separately forecast the national popular vote shares and the relative positions of the states.

Keywords: Bayesian updating, election prediction, pre-election polls, shrinkage estimation

1 Introduction

Research tells us that national elections are predictable from fundamentals (e.g., Rosenstone, 1983, Campbell, 1992, Gelman and King, 1993, Erikson and Wlezien, 2008, Hibbs, 2008), but this doesn't stop political scientists, let alone journalists, from obsessively tracking swings in the polls. The next level of sophistication—afforded us by the combination of ubiquitous telephone polling and internet dissemination of results—is to track the trends in state polls, a practice which was led in 2004 by Republican-leaning realclearpolitics.com and in 2008 at the websites election.princeton.edu (maintained by biology professor Sam Wang) and fivethirtyeight.com (maintained by Democrat, and professional baseball statistician, Nate Silver).

*We thank Aaron Strauss for helpful comments and the National Science Foundation, Yahoo Research, and the Columbia University Applied Statistics Center for partial support of this work.

Presidential elections are decided in swing states, and so it makes sense to look at state polls. On the other hand, the relative positions of the states are highly predictable from previous elections. So what is to be done? Is there a point of balance between the frenzy of daily or weekly polling on one hand, and the supine acceptance of forecasts on the other? The answer is yes, a Bayesian analysis can do partial pooling between these extremes. We use historical election results by state and campaign-season polls from 2000 and 2004 to estimate the appropriate weighting to use when combining surveys and forecasts in the 2008 campaign.

The year leading up to a presidential election is full of polls and speculation, necessitating a study of the measure of uncertainty surrounding predictions. Given the true proportion who intend to vote for a candidate, one can easily compute the variance in poll results based on the size of the sample. However, here we wish to compute the forecast uncertainty given the poll results of each state at some point before the election. To do this, we need not only the variance of a sample proportion, but an estimate for how much the true proportion varies in the months before the election, and a prior distribution for state-level voting patterns. We base our prior distribution on the 2004 election results and use these to improve our estimates and to serve as a measure of comparison for the predictive strength of pre-election polls.

We use as an example the polls conducted in February, 2008, by SurveyUSA, which sampled nearly 600 voters in each state, asking the questions, "If there were an election for President of the United States today, and the only two names on the ballot were Republican John McCain and Democrat Hillary Clinton, who would you vote for?" and "What if it was John McCain against Democrat Barack Obama?" The poll was conducted over the phone using the voice of a professional announcer, with households randomly selected using random digit dialing (Survey Sampling International, 2008). Each response was classified as one of the two candidates or undecided. For each state the undecided category consisted of 5–14% of those polled, and these people as well as third-party supporters were excluded from our analysis. Likewise, for previous election results, we restrict the population to those who supported either the Democrat or the Republican.

This paper merges prior data (the 2004 election results) and the poll data described above to give posterior distributions for the position of each state relative to the national popular vote. For the national popular vote we use a prior determined by Douglas Hibbs's "bread and peace model" (Hibbs, 2008), and again merge with our SurveyUSA poll data.

In sections 2 and 3 of this article we ascertain the strength of each source of data in

predicting the election. Section 2 contains an analysis of the use of past election results in predicting future election results, ultimately resulting in an estimate for the variance of the 2008 relative state positions given the 2004 election results. Section 3 contains an analysis of the strength of pre-election polls in predicting election results, giving measures both of poll variability and variability due to time before the election. Section 4 brings the sources together with a full Bayesian analysis, fusing prior data with poll data to create posterior distributions.

2 Past Election Results

The political positions of the states are consistent in the short term from year to year; for example, New York has strongly favored the Democrats in recent decades, Utah has been consistently Republican, and Ohio has been in the middle. We begin our analysis by quantifying the ability to predict a state outcome in a future election using the results of past elections. We do this using the presidential elections of 1976–2004. We chose not to go back beyond 1976 since state results correlate strongly ($.79 \leq r \leq .95$) for adjacent elections after 1972, while the correlation between the 1972 and 1976 elections is only .11.

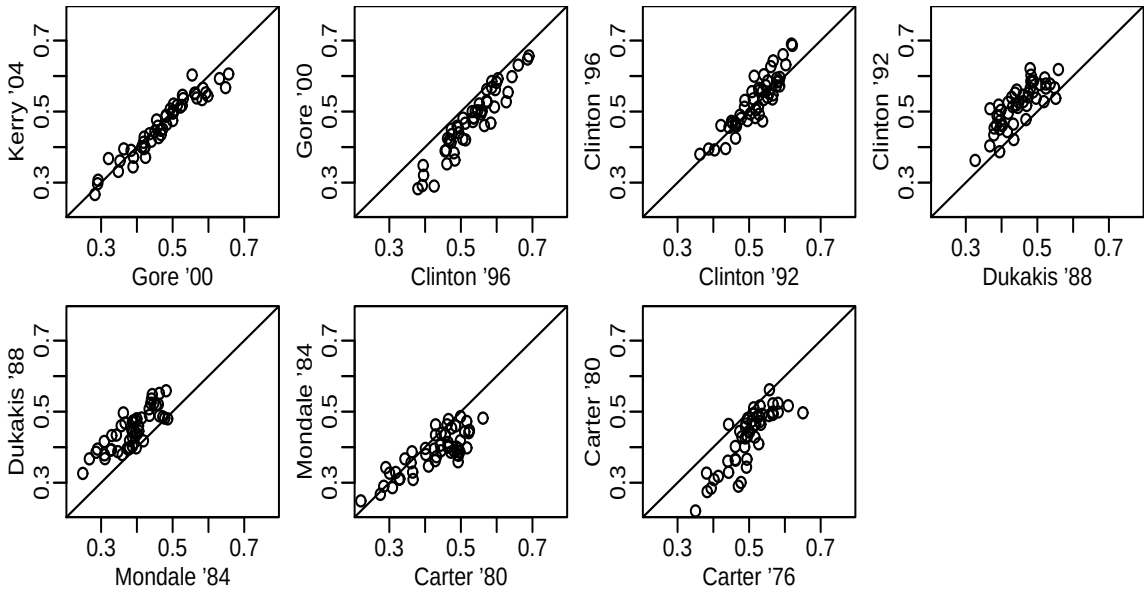


Figure 1: State results from one presidential election to the next, in each case showing the Democratic candidates' share of the two-party vote in each state.

Figure 1 shows strong correlations in the Democratic share of the vote in each state from

one presidential election to the next. But in many cases the proportion for the Democrat is uniformly higher or lower than would have been predicted by the previous election. For example, states had much higher proportions for Clinton in 1992 than for Dukakis in 1988, and much lower proportions for Gore in 2000 than for Clinton in 1996. This does not indicate a change in state's relative partisanship but rather a varying nationwide popularity of the Democratic candidate from election to election. The popularity of Kerry may not predict the popularity of Obama, but the popularity of Kerry in any given state compared to the popularity of Kerry nationwide seems to be indicative of the future popularity of Obama in that state as compared to nationwide. For this reason we look at the *relative state positions*, the difference between the proportion voting Democratic in each state and the national proportion voting Democratic.

We tried various models using past elections to predict future elections, but found that not much was gained by using data from elections prior to the most recent election. Therefore for simplicity, in our analysis of 2008 we ignore election data before 2004, and simply consider the proportion of voters in each state choosing John Kerry over George W. Bush in the 2004 election. Our only adjustment is a home-state correction: we subtract 6% (as determined via analysis of past elections) off the vote for Bush and Kerry in Texas and Massachusetts, respectively, and give the same amount in the forecast for McCain in Arizona and Clinton in New York or Obama in Illinois. Finally, Kerry's share of the two-party vote was 48.9% so our prior data become, for each state, the proportion voting for Kerry minus .489.

To determine the strength of our prior data, we need to know how much these state relative positions vary from election to election. Let $d_{s,y}$ be the relative position for state s in year y. We first estimate $var(d_{s,2008}|d_{s,2004})$ for each state by $\frac{1}{7}\sum_{i=1}^{7}(d_{s,y_{i+1}} - d_{s,y_i})^2$, where $\vec{y} = (1976, \ldots, 2004)$. With only seven data points for each state, however, these estimates could be unreliable. We could get around this problem by assuming a common variance estimate for all states, but rather than forcing either one common estimate or fifty individual estimates, we use shrinkage estimation, partial pooling. Exactly how much to pull each estimate to the common mean is determined via lmer, the tool for mixed effects models in R, and is based upon comparisons of within-state and between-state variability. Before pooling, the estimates of standard deviation for each state range from .011 to .073, with complete pooling the common estimate is .037; after our partial pooling the estimates range from .029 to .056.

From the normal approximation, we can expect the difference in 2008 to fall within .06

of the 2004 state difference for the most consistent states and up to .11 away for the least consistent states.

3 Pre-Election Polls

How much can we learn from a February poll of 600 voters in each state? If we ignore that the poll was conducted so early in the year, it appears we can learn quite a lot. Due to sampling variability alone, we would expect the true proportion who would vote Democratic in each state to be within .04 of the sample proportion (SD $= \sqrt{p(1-p)/n} \approx \sqrt{.5 \cdot .5/600} = .02$). A standard deviation of .02 would make a poll of this size more informative than the 2004 election. Using Monte Carlo techniques, one could simulate many potential "true" proportions for each state, and so many potential popular or electoral college results, as done in Erikson and Sigman (2008). However, this would depict voter preferences *in February.* To get a true measure of variability, we need to consider not only sampling variability, but variability due to time before the election.

We first estimate the variance in the national popular vote due to time before the election, using the results of Gallup polls leading up to the presidential elections of 1952 through 2004. Let p_t denote the true national proportion who would vote Democratic t months before the election, $\hat{p}_t$ denote our estimate of p_t as gotten by a pre-election poll, and p_0 denote the two-party Democratic vote share in the actual election. Ideally we'd like $var(\hat{p}_t|p_0)$ as a function of both the poll sample size, n, and the number of months before the election the poll was conducted, t. Decomposing the variance conditionally yields,

$$\begin{aligned} var(\hat{p}_t|p_0) &= E(var(\hat{p}_t|p_t)|p_0) + var(E(\hat{p}_t|p_t)|p_0) \\ &= E\left(\frac{p_t(1-p_t)}{n}|p_0\right) + var\left(p_t|p_0\right) \\ &= \frac{E(p_t|p_0) - E(p_t^2|p_0)}{n} + var\left(p_t|p_0\right) \\ &= \frac{p_0(1-p_0)}{n} + \left(\frac{n-1}{n}\right) var\left(p_t|p_0\right) \\ &\approx \frac{p_0(1-p_0)}{n} + var\left(p_t|p_0\right). \end{aligned} \tag{1}$$

Thus $var(p_t|p_0) = var(\hat{p}_t|p_0) - p_0(1-p_0)/n$, so can be estimated by empirically calculating $var(\hat{p}_t|p_0)$ and subtracting off the expected sampling variability. Let $\hat{p}_{t,i}$ and $n_{t,i}$ denote estimated proportion and sample size respectively for the i^{th} poll in a given month, and let N_t be the number of polls we have t months before the election (from Gallup polls

1952 to 2004). We then estimate $var(p_t|p_0)$ by

$$\widehat{var}(p_t|p_0) = \frac{\sum_{i=1}^{N_t}\left[(\hat{p}_{t,i} - p_0)^2 - \frac{p_0(1-p_0)}{n_{t,i}}\right]}{N_t}. \quad (2)$$

The standard deviations estimated in this fashion for each month are displayed in Figure 2(a). We then fit a linear regression to these points, with an intercept of 0 (we are assuming the popular vote in November should match that of the election and ignoring issues such as voter turnout). This model gives $\widehat{SD}(p_t|p_0) = .012t$, with a standard error of .0012 on the slope, suggesting that the standard deviation in the underlying popular vote increases by .012 each additional month before the election. This estimates $\widehat{SD}(p_{feb}|p_0) = .11$, essentially saying that February polls contain almost no information about the popular vote at the time of the election.

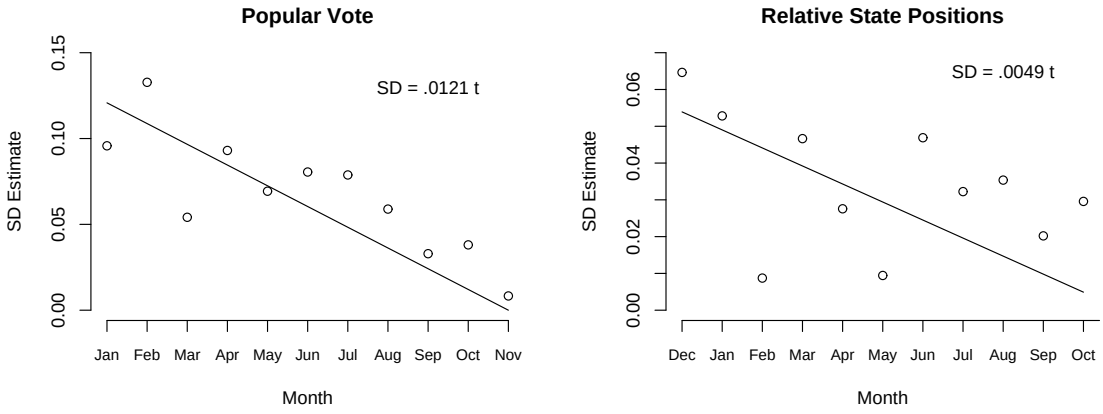


Figure 2: (a) Estimated standard deviations of the popular vote in each month given the popular vote in the election. (b) Estimated standard deviations of the relative position of each state in each month, given the relative position of the state in the election.

We now perform the same essential calculations as above, but for the variance of the relative state positions due to time before the election. In both 2000 and 2004, the Annenberg Public Policy Center at the University of Pennsylvania conducted the National Annenberg Election Survey (NAES), a series of polls throughout the year leading up to the election. Again restricting our analysis only to those who say they would vote for the Democrat or the Republican, we have 43,373 people polled in 2000 and 52,825 in 2004.

Now we want $var(\hat{d}_{s,t}|d_0)$ as a function of n and t, where $d_{s,t}$ is the relative position of state s, t months before the election. We follow the same logic as with the popular vote, except now instead of averaging over multiple years worth of pre-election polling data, with only two years to work with we have to average over the states (assuming a common

variance for all states). For each state, each month, sample sizes range from 0 to 844, but with 42% having less than 30 people polled. Sample sizes this small lead to unreliable estimates, so we tweak (2) slightly and take a weighted average, weighting by sample size. We thus estimate $var(d_{s,t}|d_0)$ by

$$\widehat{var}(d_{s,t}|p_{s,0}) = \frac{\sum_{y\in\{2000,2004\}}\sum_{s=1}^{50} n_{s,y,t}\left[(\hat{d}_{s,y,t} - d_{s,y,0})^2 - \frac{p_{s,y,0}(1-p_{s,y,0})}{n_{s,y,t}}\right]}{\sum_{y\in\{2000,2004\}}\sum_{s=1}^{50} n_{s,y,t}}. \quad (3)$$

This isn't quite as straightforward as the calculation for (2), since we don't observe the national popular vote at time t so can't actually observe $\hat{d}_{s,t}$ (we only have $\hat{p}_{s,t}$). To get around this we need an estimate for the popular vote each month before the elections of 2000 and 2004. The election outcome, the Annenberg state polls, and Gallup poll data each give us an estimate of the popular vote for each month. The strength of the poll data will depend on the sample size for that particular month, and the strength of the actual election popular vote will depend on how many months before the election we are trying to estimate, so our estimate for January will be almost all poll-based, while our estimate for November will be entirely based on the election outcome. Luckily, we just developed a formula for $var(p_t|p_0)$ which we can use again here to determine how much to weight the election outcome for each month. We estimate the popular vote for each month by weighting the estimates from the election, Annenberg polls, and Gallup polls each by their respective information. The standard deviations for these weighted estimates range from .003 in months right before the actual election where the election results are very informative and many polls are conducted, up to .014 for earlier months without a lot of poll data.

We use these popular vote estimates to calculate each $\hat{d}_{s,t}$, which then allows us to compute (3) for each month. The estimated standard deviations are shown in Figure 2(b). The linear regression fit to these data points, again with intercept 0, gives the equation $\widehat{SD}(d_{s,t}|d_{s,0}) = .0049t$, with a slope standard error of .00085. This estimates $\widehat{SD}(d_{s,feb}|d_{s,0}) = .044$.

4 Posterior Distributions

With the variance estimates derived in sections 2 and 3, we are all set to go forth with the full Bayesian analysis. We first look only at the relative positions of the states, and momentarily ignore the national popular vote.

For our poll data, we look at $\hat{d}_{s,feb}$ for each state. We don't know the popular vote in February so can't compute these exactly, but can get a pretty close estimate given that

we have a sample size exceeding 500 in each state. The relative positions based on our February poll data given the relative positions in the election follow a normal distribution (a reasonable approximation given the large sample size in each state), with variance incorporating both sampling variability and our estimate of variance due to the polls being conducted in February (section 3):

$$\hat{d}_{s,feb}|d_{s,0} \sim N\left(d_{s,0}, \frac{p_{s,0}(1-p_{s,0})}{n_{s,feb}} + .0441^2\right). \tag{4}$$

The sample sizes range from 500 to 600, leading to standard deviations ranging from .055 to .057.

For our prior, the 2004 election data, we have

$$d_{s,2008}|d_{s,2004} \sim N\left(d_{s,2004}, var(d_{s,2008}|d_{s,2004})\right). \tag{5}$$

Recall from Section 2 that $var(d_{s,2008}|d_{s,2004})$ varies from state to state, and ranges from $.029^2$ to $.056^2$. For almost all states this standard deviation is smaller than that of the poll data, meaning our posteriors will usually be closer to the 2004 election results than to the February polls.

We use a normal-normal mixture model to create the posterior, weighting by information, the reciprocal of variance. This gives $d_{s,2008}|(d_{s,2004}, \hat{d}_{s,feb}) \sim$

$$N\left(\frac{\left(\frac{1}{var(\hat{d}_{s,feb}|d_{s,0})}\right)\hat{d}_{s,feb} + \left(\frac{1}{var(d_{s,0})}\right)d_{s,2004}}{\frac{1}{var(\hat{d}_{s,feb}|d_{s,0})} + \frac{1}{var(d_{s,0})}}, \quad \frac{1}{\frac{1}{var(\hat{d}_{s,feb}|d_{s,0})} + \frac{1}{var(d_{s,0})}}\right). \tag{6}$$

For a typical state, this simplifies to something like

$$d_{s,2008}|\hat{d}_{s,feb}, d_{s,2004} \sim N(.35\hat{d}_{s,feb} + .65d_{s,2004}, .03^2),$$

with the weight on the poll estimate ranging from .26 to .56 and the standard deviations ranging from .025 to .037, and with higher standard deviations for states with more weight on the polls. Figure 3 shows the posterior predictions for the relative positions of the states for both Clinton and Obama. (The poll was conducted before the Democratic candidate was chosen, and our prior applies to any Democratic candidate.)

We now move on to creating a posterior for the national popular vote. We construct our prior based on the estimate and predictive standard deviation from Hibbs (2008), who predicts the national two-party Democratic vote share based only on two factors: weighted-average growth of per capita real personal disposable income over the previous term, and

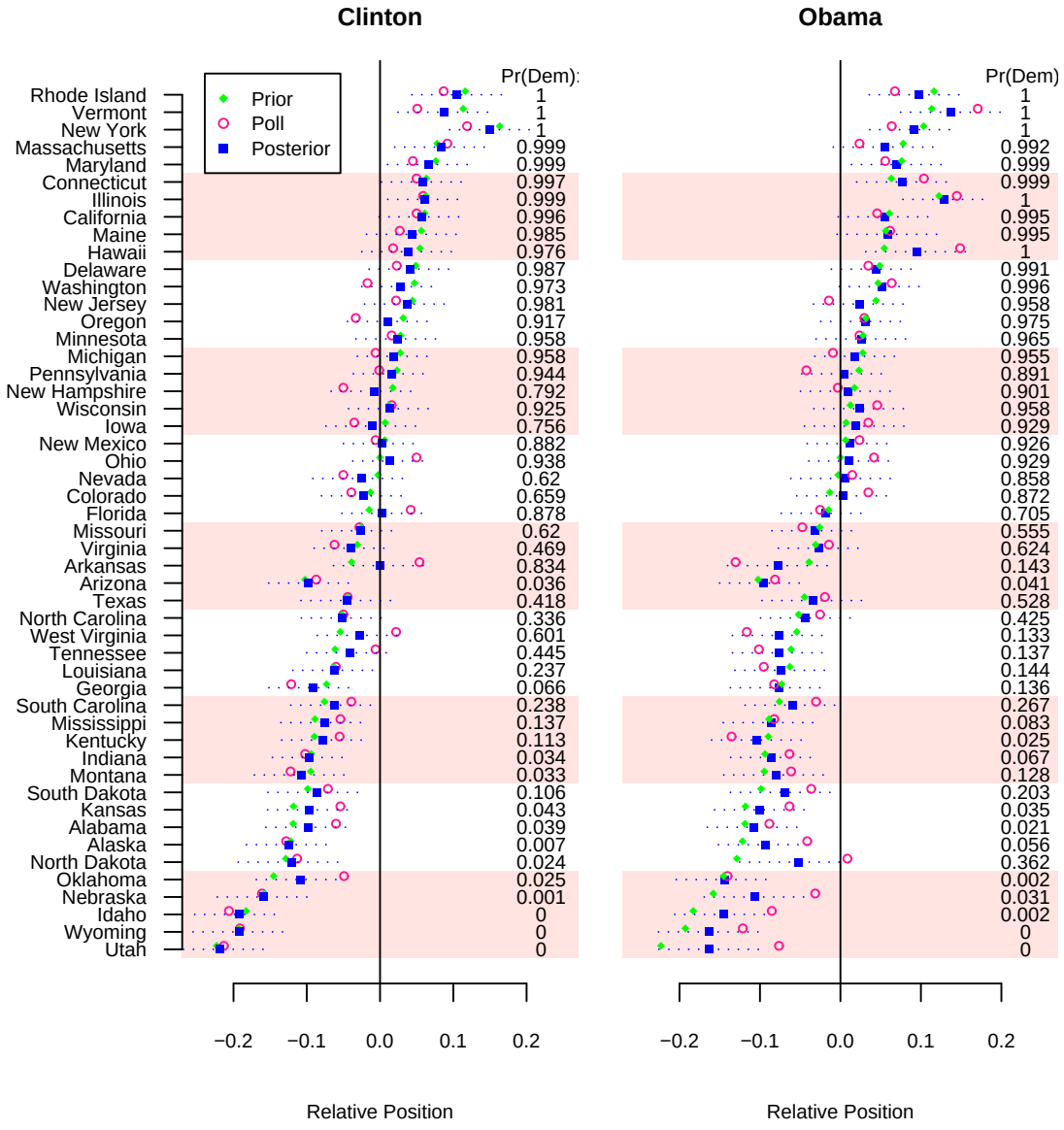


Figure 3: 95% posterior intervals for the relative position of each state, alongside prior and poll point estimates. The left column gives the probability of each state going Democratic (which incorporates the posterior for the national popular vote). States are ordered by 2004 Democratic vote share.

cumulative US military fatalities owing to unprovoked hostile deployments of American armed forces in foreign conflicts. To determine the variance in the success of this model we look at its predictions for the last 14 elections (1952 to 2004). The sample standard deviation of (predicted − actual) is .0208 (quite accurate for only two predictors and no polling information!). Shortly before the election, Hibbs predicted that Obama would get 53.75% of the two-party vote.

With our February poll data we weight the sample poll proportion voting Democratic in each state by the number of voters in that state in the 2004 election, and get a national estimate of 51.43% for Obama. From section 3, $var(\hat{p}_{feb}|p_0) = p_0(1-p_0)/n + .109^2 \approx (.5 \cdot .5)/27000 + .109^2$, giving a standard deviation of eleven percentage points. This variance may not be entirely accurate since the variance was estimated in section 3 using polls of a nationwide sample rather than a sample within each state, but we didn't have sufficient state level data from enough past elections to provide a better estimate. This estimate (.109) is much larger than the standard deviation associated with our prior (.0208), so the posterior will be strongly weighted towards Hibbs's estimate.

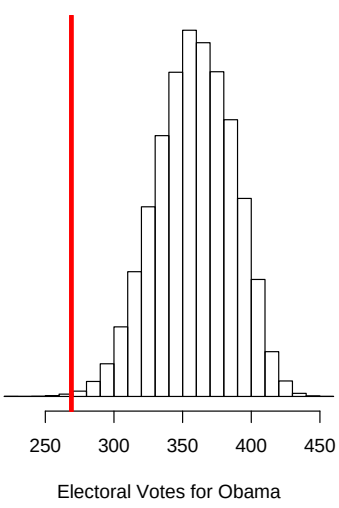
Our posterior distribution for the national popular vote, again using a normal-normal mixture model, is

$$\begin{aligned} p_{2008}|\hat{p}_{feb}, \hat{p}_{hibbs} &\sim N\left(.035\hat{p}_{feb} + .965\hat{p}_{hibbs}, \frac{1}{(1/.109)^2 + (1/.021)^2}\right) && (7) \\ &\sim N\left(.537, .020^2\right). && (8) \end{aligned}$$

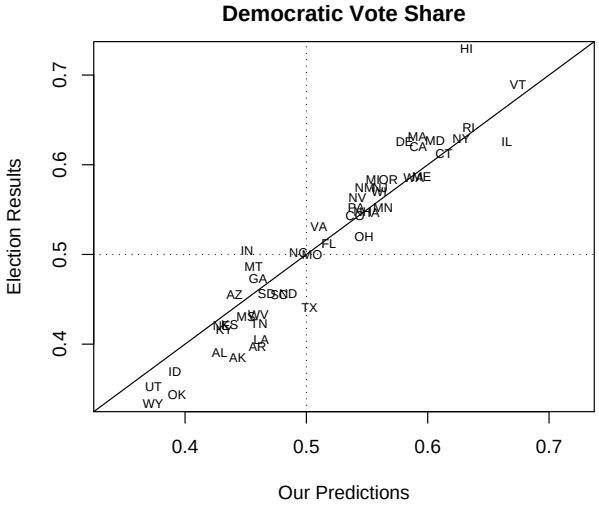
Now that we have posterior distributions for both the national popular vote and each state's position relative to this, we can simply add them together to get posterior distributions for the proportion voting Democratic in each state. To create a posterior distribution for Obama's electoral college vote share, we simulate 100,000 elections, each time randomly drawing first a national popular vote from (8), and then simulating each state outcome by adding a draw from (6) to the simulated popular vote. The simulated electoral vote outcomes are shown in Figure 4(a), and have a posterior mean of 359 and SD of 28. Of the 100,000 simulated elections, Obama won 99,886.

5 Conclusion

Our predictions were based on the SurveyUSA February poll data (for both the relative state positions and the popular vote estimate), the 2004 election results (for the relative state positions), and Hibbs' October estimate of the popular vote. We ignore the past 8 months'



(a) Posterior distribution for Obama's electoral college vote share. Anything $\geq$ 270 indicates an Obama victory.



(b) Our predicted proportion voting Democratic in each state versus the actual election results.

furor of pre-election polling (March to October), and any effect of either candidate's campaign has absolutely no impact on our prediction. Our analysis and paper up to this point were completed in entirety before November, 2008, yet this paragraph is added just after the election, allowing us to compare our posterior estimates with the actual election results. The actual two-party popular vote for Obama was 53.4%, while our posterior prediction was 53.7%. Figure 4(b) shows our predicted Democratic vote share for each state against the actual results. One can see that while we came quite close for most states, we tended to overestimate Obama's popularity in Rebublican states and underestimate in Democratic states. The correlation between our predicted values and actual values is .96, and the root mean square error (RMSE) of our estimates is $\sqrt{(1/50)\sum_{s=1}^{50}(p_{s,predicted} - p_{s,actual})^2} = .032$. The RMSE for fivethirtyeight.com's estimates, which use polls leading up the election, is .025. It is not surprising that you get closer to the truth using pre-election polls right before the election, but it is remarkable that we can do so well without using any polling data collected beyond February.

While the accuracy of our predictions is important, we also care about the accuracy of our variance estimates, as every prediction needs an accompanying degree of uncertainty. The RMSE for our estimated relative state positions as compared to the election results is .031, while our average posterior standard deviation is .029. The closeness of these two numbers may help to improve the credibility of our variance estimates. Across states it

appears our posterior intervals were close to the correct widths, as the true relative position of each state falls within our 95% posterior intervals for 48 of the 50 states (we underestimated Hawaii and Indiana), giving 96% coverage. (Some of this has to be attributable to luck—the state estimates are correlated, and a large national swing could easily introduce a higher state-by-state error rate.)

This paper has the goal of determining the strength of past elections and of pre-election polls in predicting a future election, and combining these sources to forecast the election. We found that to predict the current election, using the results of the most recent election is a good predictor of the way each state votes compared to the nation, but not necessarily of the national vote.

Hence, past election data are best used with a current estimate of the popular vote (such as can be obtained from polls or from forecasts that use economic and other information). Thus, our key contribution here is to separate the national forecast (on which much effort has been expended by many researchers) from the relative positions of the states (for which past elections and current polls can be combined in order to make inferences). Pre-election polls, not surprisingly, are more reliable as they get closer to the election. Our advance with this analysis is quantification of this trend.

References

[1] Annenberg Public Policy Center (2008). `www.annenbergpublicpolicycenter.org/AreaDetails.aspx?myId=1`, June.

[2] Campbell, J. E. (1992). "Forecasting the Presidential Vote in the States," *American Journal of Political Science*, 36: 386-407.

[3] Erikson, R. S., and Sigman, K. (2008). "Guest Pollster: The Survey USA 50 State Poll and the Electoral College," Pollster.com, `www.pollster.com/blogs/guest_pollster_the_surveyusa_5.php`, March.

[4] Hibbs, D. A. (2008). "Implications of the "Bread and Peace" Model for the 2008 US Presidential Election," *Public Choice*, September.

[5] Gelman, A. and King, G. (1993) "Why are American Presidential Election Campaign Polls so Variable When Votes are so Predictable?" *British Journal of Political Science*, 23: 409-451.

[6] Rosenstone, S. J. (1983). *Forecasting Presidential Elections*, New Haven, Conn.: Yale University Press.

[7] Silver, N. (2008). `www.fivethirtyeight.com/`, August.

[8] Strauss, A. (2007). "Florida or Ohio? Forecasting Presidential State Outcomes Using Reverse Random Walks," Princeton University Political Methodology Seminar.

[9] Survey Sampling International (2008). `www.surveysampling.com`, June.

[10] Wlezien, C., and Erikson, R. S. (2007). "The Horse Race: What Polls Reveal as the Election Campaign Unfolds," *International Journal of Public Opinion Research*, 19: 74-88.

[11] Wlezien, C., and Erikson, R. S. (2004). "The Fundamentals, the Polls, and the Presidential Vote," *Political Science and Politics*, 37: 747-751.