Presidential election forecast

Carl Ehrett

7 November 2016

Abstract

In addition to the interest inherent in gaining insight into the likely outcome of the 2016 U.S. presidential election, it is desirable to predict the outcome of that race in each state, so that last-minute campaign resources can be allocated accordingly. Potential sources of evidence to ground such predictions include polls assessing voter intentions and historical data about outcomes of previous presidential elections in each state. In this paper I present a model based on these two sources. The model views the election in each state i as an outcome of a bernoulli p_i trial. A set of (roughly) weekly polls in each state is also treated each as an outcome of a p_i trial. To accommodate shifts in p_i throughout the course of the campaign, these polls are included by way of a power prior with exponent a_0. Electoral history in each state is used to form an informative prior on p_i. Using Metropolis-Hastings-within-Gibbs Markov-chain Monte Carlo sampling, values for p_i and a_0 are then estimated. The values for p_i are used to draw an outcome for the election. I conclude by predicting that Hillary Rodham Clinton will succeed in becoming the 45$^{\text{th}}$ president of the United States of America.

1 Introduction

There are a wealth of different potential predictors that may be included in an attempt to predict the outcome of the 2016 presidential election. This election is widely perceived to be exceptional in various ways, and this perceived uniqueness can provide temptation to draw on a diverse array of such predictors. For example, given that there appear to be powerful and interesting demographic splits between the two major-party candidates[1], one might justifiably attempt to predict the rates of voter turnout amongst various demographic groups, hoping thereby to better predict the behavior of voters overall. Less justifiably, one may attempt to prognosticate from the basis of a purported insight into the American soul—e.g., that the depth of popular anger against Washington combined with racism and sexism amongst the electorate will propel Trump to power[2], or that a popular desire to have at last our first female president will win the day for Clinton[3].

Though there is likely some predictive hay to be made out of some such perspectives, any approach which treats this election as exceptional is inherently vulnerable both as to potential mischaracterization of its supposed exceptionality and to a lack of data and precedent for appropriately handling any genuinely exceptional features.

[1] c.f. Alcantara, Uhrmacher, & Guskin, 2016
[2] Moore, 2016
[3] Seib, 2016

Therefore this paper eschews any attempt to capture any sense in which this election is exceptional in comparison with other U.S. presidential elections. Instead, the sources of data for this paper are twofold: previous electoral history, and polls directly asking voters to self-report their electoral intentions.

Because polls are not perfectly analogous to an election, and because voter preferences shift over the course of a campaign cycle, a power prior is employed to incorporate polls into an analysis of the Bernoulli success probability p_i of state i casting its electoral votes for Hillary Clinton. The electoral history (form the past 40 years) is used to generate an informative prior for p_i. Using MCMC, the power prior exponent a_0 is sampled via a Metropolis-Hastings step; Gibbs sampling is used to draw p_i in each step (for each state i).

The methodology employed in this paper is discussed in section 2. Details about the data used in this analysis are given in section 3. The results of this analysis, including the prediction that Hillary Rodham Clinton will earn the title of president, are presented in section 4. `R` code used for this analysis is included as an appendix.

2 Methodology

For each state i (where the District of Columbia is included as a state), the electoral history of state i is used to generate an informative beta prior for p_i, the Bernoulli success probability of Clinton victory in that state. The initial prior for each such p_i is set as beta(2,2). The hyperparameters of (shape, rate) = (2,2) are chosen to indicate mild confidence that p_i is unlikely to be very near 0 or 1. The outcome in state i in each previous election (from the past 40 years) is seen as a Bernoulli(p_i) trial (where Democratic victory is coded as success). Thus the sum of Democratic victories in presidential elections in state i is a Binomial$(10, p_i)$. Where E_i is the number of Democratic victories in state i, then, we have that the informative prior for this election takes p_i to be distributed as a beta$(E_i + 2, 10 - E_i + 2)$; i.e., beta$(E_i + 2, 12 - E_i)$.

The j^{th} poll in state i from this election is treated as a realization from a binomial(n_{ij}, p_i), where n_{ij} is the number of polled likely voters[4] in the j^{th} poll of state i. To accommodate both shifting voter preferences during the election cycle and possible differences between polling responses and election behavior, a power prior is used. Specifically: the election itself is treated as the final "poll". For $j \neq 1$, the posterior distribution of p_i given the j^{th} poll relies on a power prior based on the $j - 1^{\text{st}}$ poll. Since the posterior of p_i given the $j - 1^{\text{st}}$ poll itself relies on a power prior based on the $j - s^{\text{nd}}$ poll (for $j > 2$), the result is that the posterior distribution of p_i given the final poll (just prior to the election) implicitly is influenced by all previous polls by way of decreasing power prior exponents. Thus polls closer to the election have a proportionally greater impact on the distribution of p_i than do earlier polls. The power prior exponent a_0 is treated as unknown with a uniform(0,1) prior.

The resulting model can be expressed as follows:

[4]See section 3 for more discussion of which poll responses are used.

Definitions:	
E_i	Total number of Democratic victories in last 40 years in state i
m	Total number of polls
p_i	Bernoulli probability of Democratic victory in state i
p	Vector of p_i for $i = 1, \cdots, 51$
D_{ij}	Ordered pair (n_{ij}, Y_{ij})
n_{ij}	Total poll responses from poll j in state i
Y_{ij}	Total pro-Clinton responses from poll j in state i
D_i	Matrix containing D_{ij} for $j = 1, \cdots, 10$
D	Matrix containing D_i for $i = 1, \cdots, 51$
a_0	Power prior exponent for effect of polls on p_i

Informative prior:

$$\pi(p_i|E_i) \sim \text{beta}(2 + E_i, 12 - E_i)$$

Posterior:

$$\pi(p_i|D_{im}) \sim \text{beta}\left(2 + E_i + \sum_{k=1}^{m} Y_{ik} \cdot a_0^{m-k}, 12 - E_i + \sum_{k=1}^{m}(n_{ik} - Y_{ik})a_0^{m-k}\right)$$

The posterior distribution of a_0 is not easy to draw directly. Therefore, this paper employs the Metropolis-Hastings algorithm to sample a_0, whereas Gibbs sampling is employed to draw $p_i|D_{im}$. Since a_0 is constrained to be in $[0, 1]$, the logit transform is used to eliminate these boundary constraints. That is: for a given a_0, define $g_0 = \log\left(\frac{a_0}{1-a_0}\right)$. To propose a new candidate a_0^*, first draw g_0^* from a normal distribution centered at g_0 with variance 1. Note that it would be simple to employ an adaptive burn-in period in which the variance of the proposal distribution is attenuated to achieve an optimal acceptance rate. However, the variance of 1 is successful enough that such an adaptive technique is not needed here (for more details on this, see section 4). Next, let $a_0^* = \frac{e^{g_0^*}}{1+e^{g_0^*}}$. Then where f is the proposal distribution (in terms of a_0 rather than g_0), find the Metropolis-Hastings acceptance ratio a:

$$a = \min\left\{1, \frac{\pi(p|D, a_0^*) \cdot f(a_0|a_0^*)}{\pi(p|D, a_0) \cdot f(a_0^*|a_0)}\right\}$$

In this paper, we assume $p_i \perp\!\!\!\perp p_j$ for $i \neq j$. This assumption is made for simplicity. However, it is likely not strictly true. Therefore, an improvement on the analysis presented here could involve allowing for correlation amongst the states.

Note that the reason that Metropolis-Hastings is required here (rather than Metropolis) is that the proposal distribution f is not symmetric. Although the normal distribution is symmetric, we are working here with a normal distribution amongst logit transforms of the a_0, and thus symmetry is lost. The ratio of the two proposal pdfs reduces to the ratio of the derivative of the logit transform; that is:

$$\begin{aligned}\frac{f(a_0|a_0^*)}{f(a_0^*|a_0)} &= \frac{1/[a_0(1-a_0)]}{1/[a_0^*(1-a_0^*)]} \\ &= \frac{a_0^*(1-a_0^*)}{a_0(1-a_0)}\end{aligned}$$

Thus, we have:

$$a = \min\left\{1, \frac{\prod_{i=1}^{51} \pi(p_i|D_i, a_0^*) \cdot a_0^*(1-a_0^*)}{\prod_{i=1}^{51} \pi(p_i|D_i, a_0) \cdot a_0(1-a_0)}\right\}$$

Thus the MCMC algorithm proceeds by, for each sample, drawing p_i from $pi(p_i|D_im)$ for all i, and then by drawing a_0 using the Metropolis-Hastings technique described above. Each p_i is then used to perform a Bernoulli draw, to find whether state i votes Democratic on this "round". The results of these 51 Bernoulli draws are then used, along with information about how many electoral votes each state possesses, to determine whether Clinton or Trump has won the presidency[5].

Note that this approach treats Nebraska and Maine each as unified electoral voting blocks. In fact, each of these states is divided into two independent electoral voting blocks. The polls used in this paper identify respondents, however, only by state. Therefore, the use of this particular set of polling data requires the simplifying assumption that Nebraska and Maine behave in a unified fashion. An improved version of the analysis in this paper would acquire polling data that distinguishes intra-state voting blocks.

The resulting MCMC algorithm to draw M samples may be expressed as follows:
For $k = 1, \cdots, M$:

1. Use Metropolis-Hastings step described above to draw $a_0^{(k)}$.
2. Use posterior distribution of p to draw $p^{(k)}$.
3. Draw Bernoulli(p_i) for $i = 1, \cdots, 51$ to get election results for each state.
4. Use information on electoral votes for each state to determine national election result.

3 Data

Information about the electoral outcomes by state in the past 10 election cycles was taken from the website of the Office of the Federal Register[6]. This paper uses only the past 10 elections for two major reasons. One is that the two major parties have evolved significantly since the Civil Rights era, and thus going farther back than the mid-1970's is potentially distorting. Secondly, and more importantly, the modern primary system was not fully developed until the early 1970's. This fact makes our current election much more analogous, at least procedurally, to those taking place from 1976 onward to those occuring prior to the 1970's.

The same website provides information about how many electoral votes each state may cast.

The polls used in this paper are Google Consumer surveys made public by Google. These polls were chosen primarily because they are carried out at regular intervals (roughly once per week) in every state simultaneously, and also because they include information relevant to determining the respondent's status as a likely voter. No national polls were used in this paper. An improvement on this paper would expand it to include a greater diversity of polls, including national polls.

Google Consumer surveys are presented as a "surveywall" [7]. Similar to a paywall which may confront users seeking premium content, a surveywall is a popup which may appear on certain websites, requiring the user to complete the survey in order to access their desired content. The complete survey appears as figure 1 (although the order of the candidates is randomized each time the survey is offered). Notice

[5]Note that this ignores the (small but real) possibility that Evan McMullin will win Utah. An improved version of the analysis in this paper would accommodate the possibility of states being awarded to more than two candidates.

[6]https://www.archives.gov/federal-register/electoral-college/votes/index.html

[7]McDonald, Mahebbi, & Slatkin, 2015

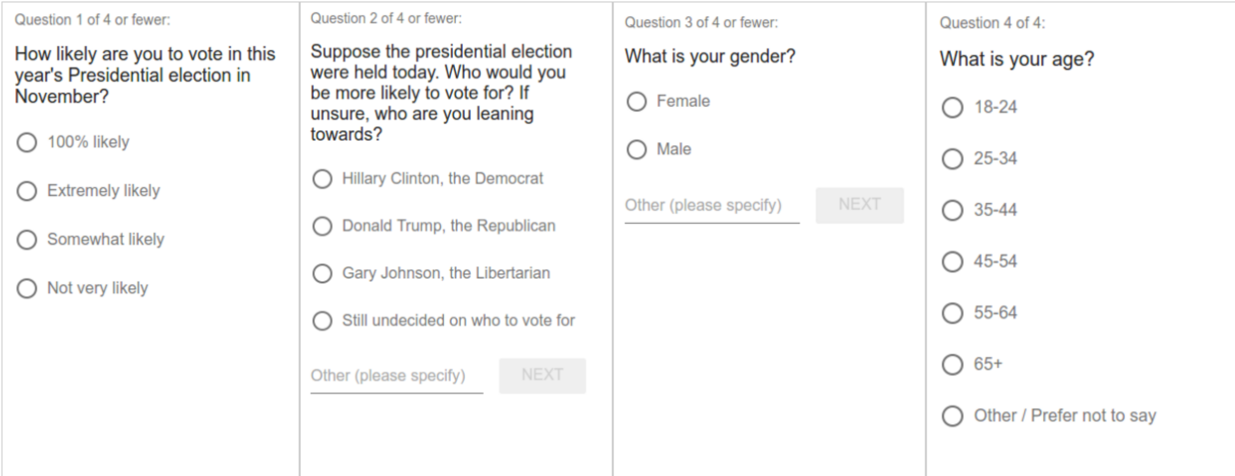

Figure 1: Google Consumer Surveys election poll.

that the survey does not ask users to self-report location, as this is inferred from their IP address. User responses are used to dynamically target a representative age $\times$ gender $\times$ state distribution of users.

In this paper, only likely voters were considered. Likely voters were identified as those respondents having identified themselves as either "100% likely" or "Extremely likely" to vote in the election. Third-party and undecided voters were not considered; thus, only decided, two-party likely voters are included in this analysis. An improved version of this analysis would attempt to model the behavior of undecided voters, and (probably less importantly) that of third-party voters. It would also be possible to develop a more nuanced method of identifying likely voters than that used here.

4 Forecast

The model and data used in this analysis predict that Hillary Rodham Clinton will win the election. This is based on a sample of $M = 100,000$ with 5000 samples discarded as burn-in.

The model does not favor Clinton terribly strongly, however. Of the samples drawn, the proportion in which she wins is roughly 0.54. The results in each state are displayed below. The victor is identified via examining the mean of the p_i samples. This is presented along with an $\alpha = 0.1$ HPD interval for each p_i. (Recall that p_i is the probability of Clinton winning state i.) Where 0.5 falls within the HPD interval for a state, that state is identified as a swing state.

	STATE	VICTOR	p_i	SWING	HPD INTERVAL
[1]	US-AK	Trump	0.439	Swing	0.366 0.510
[1]	US-AL	Trump	0.272		0.243 0.300
[1]	US-AR	Trump	0.448		0.402 0.494
[1]	US-AZ	Clinton	0.528		0.501 0.555
[1]	US-CA	Clinton	0.631		0.612 0.648
[1]	US-CO	Clinton	0.590		0.563 0.618
[1]	US-CT	Clinton	0.556		0.521 0.592
[1]	US-DC	Clinton	0.765		0.700 0.830
[1]	US-DE	Clinton	0.527	Swing	0.464 0.590

```
[1] US-FL   Trump     0.453          0.433 0.472
[1] US-GA   Trump     0.429          0.401 0.458
[1] US-HI   Clinton   0.663          0.597 0.729
[1] US-IA   Clinton   0.536          0.502 0.570
[1] US-ID   Trump     0.351          0.301 0.403
[1] US-IL   Clinton   0.568          0.547 0.589
[1] US-IN   Trump     0.499   Swing  0.473 0.525
[1] US-KS   Clinton   0.539   Swing  0.499 0.577
[1] US-KY   Trump     0.426          0.391 0.461
[1] US-LA   Trump     0.364          0.328 0.401
[1] US-MA   Clinton   0.653          0.623 0.683
[1] US-MD   Clinton   0.638          0.608 0.669
[1] US-ME   Clinton   0.594          0.529 0.658
[1] US-MI   Clinton   0.517   Swing  0.490 0.544
[1] US-MN   Clinton   0.553          0.525 0.581
[1] US-MO   Trump     0.492   Swing  0.466 0.519
[1] US-MS   Trump     0.334          0.295 0.374
[1] US-MT   Trump     0.420          0.369 0.473
[1] US-NC   Trump     0.461          0.435 0.487
[1] US-ND   Trump     0.313          0.246 0.378
[1] US-NE   Trump     0.375          0.330 0.418
[1] US-NH   Clinton   0.509   Swing  0.434 0.583
[1] US-NJ   Clinton   0.550          0.526 0.573
[1] US-NM   Clinton   0.584          0.540 0.628
[1] US-NV   Clinton   0.523   Swing  0.478 0.566
[1] US-NY   Clinton   0.584          0.564 0.603
[1] US-OH   Trump     0.471          0.448 0.494
[1] US-OK   Trump     0.370          0.329 0.410
[1] US-OR   Clinton   0.600          0.568 0.634
[1] US-PA   Trump     0.486   Swing  0.464 0.508
[1] US-RI   Clinton   0.595          0.519 0.674
[1] US-SC   Trump     0.419          0.384 0.454
[1] US-SD   Trump     0.333          0.270 0.395
[1] US-TN   Trump     0.397          0.365 0.430
[1] US-TX   Trump     0.413          0.393 0.432
[1] US-UT   Trump     0.487   Swing  0.438 0.535
[1] US-VA   Clinton   0.537          0.510 0.564
[1] US-VT   Clinton   0.741          0.681 0.801
[1] US-WA   Clinton   0.615          0.588 0.642
[1] US-WI   Clinton   0.588          0.564 0.613
[1] US-WV   Trump     0.379          0.322 0.435
[1] US-WY   Trump     0.286          0.224 0.347
```

To check the validity of this analysis, it is helpful to examine the autocorrelations and effective sample sizes of its national election forecast, its estimate of p_i for each i, and its estimate of a_0. Firstly, the binary predictor of national election outcome has essentially no autocorrelation: see figure 2. Accordingly, it has full effective sample size of 95,000 in a sample of 100,000 with 5,000 discarded as burn-in.

The autocorrelation of the p_i are similar; there is very little autocorrelation for any of these, and the effective sample size is near to or equal to 95,000 in all cases. The lowest effective sample size is for South Dakota, at 69607.72, which is still quite high enough for our purposes. Rather than examine plots of each of the p_i

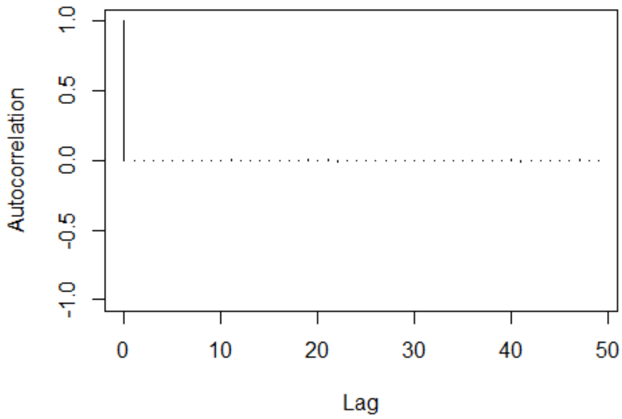

Figure 2: Autocorrelation of binary national election outcome.

and their autocorrelation, a single representative in each case is given by figures 3 and 4.

Finally, turning to a_0, we see that it suffered from far more autocorrelation than any p_i, but not from a terribly troubling amount. The effective sample size from our $M = 100,000$ draws is 8001.96. The autocorrelation may be seen in figure 5. The mean value of a_0 is approximately 0.5001. An $\alpha = 0.05$ HPD interval is given below.

```
> HPDinterval(a0.mcmc,prob=0.95)
lower     upper
var1 0.4010063 0.5936153
attr(,"Probability")
[1] 0.95
```

It appears, then, that the MCMC samples are sufficiently uncorrelated to ground reasonable confidence in the results of the analysis.

References

[1] Alcantara, C., Uhrmacher, K., & Guskin, E. (2016, October 16). Clinton and Trump?s demographic tug of war. Retrieved November 7, 2016, from `https://www.washingtonpost.com/graphics/politics/2016-election/the-demographic-groups-fueling-the-election/`

[2] Google Surveys 2016 US Election Poll. (n.d.). Retrieved November 7, 2016, from `https://drive.google.com/drive/folders/0B29GVb5ISrT0VzFQWjFSWGcyeVE`

[3] McDonald, P., Mahebbi, M., & Slatkin, B. (2015, June). Comparing Google Consumer Surveys to Existing Probability and Non-Probability Based Internet Surveys. Retrieved November 7, 2016, from `http://services.google.com/fh/files/misc/consumer_surveys_whitepaper_v2_updated.pdf`

[4] Moore, M. (2016, July 21). 5 Reasons Why Trump Will Win. Retrieved November 07, 2016, from `http://michaelmoore.com/trumpwillwin/`

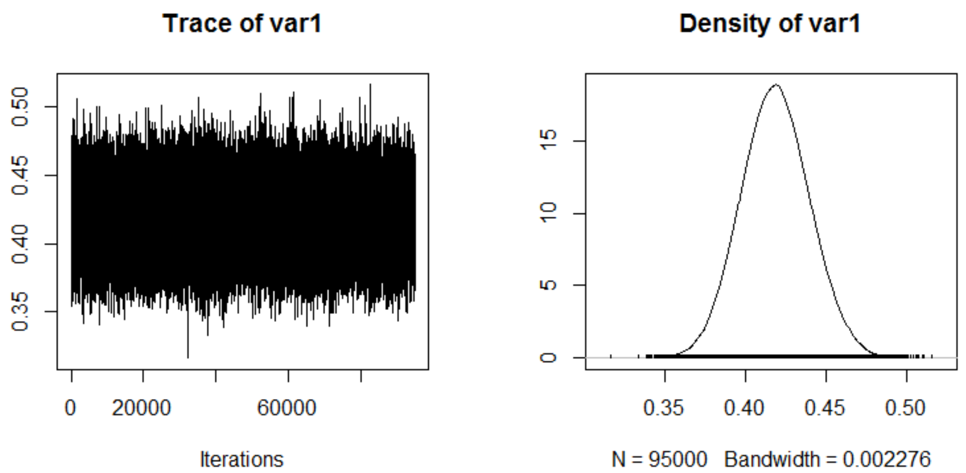

Figure 3: Plot and variance of the samples of p_i, where i is South Carolina.

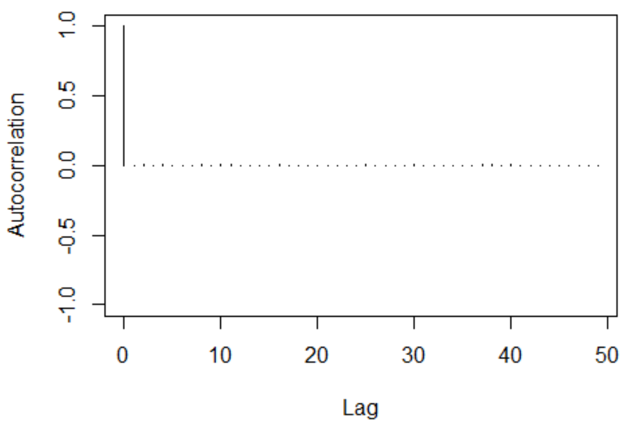

Figure 4: Autocorrelation of p_i, where i is California.

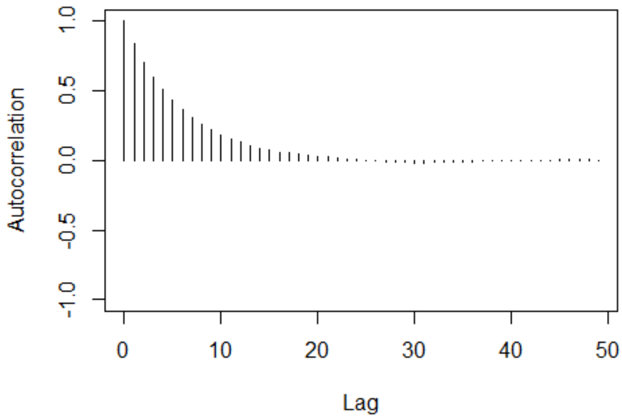

Figure 5: Autocorrelation of a_0.

[5] Seib, P. (2016, July 30). Why Clinton Will Win. Retrieved November 07, 2016, from `http://www.huffingtonpost.com/philip-seib/why-clinton-will-win_b_11279386.html`

[6] U. S. Electoral College: Historical Election Results. (n.d.). Retrieved November 07, 2016, from `https://www.archives.gov/federal-register/electoral-college/historical.html`