Predicting the 2008 U.S. Presidential Election

Benjamin Schak

Jun. 22nd 2008

1 Goal

The primary goal of this work is to answer the question, "Who will win the 2008 presidential election between Senators John McCain (R–AZ) and Barack Obama (D–IL)?"

Of course there are many other interesting questions that naturally arise, and a good framework for answering the primary question ought to suffice to answer these questions as well. For example:

- What chances do the candidates have of winning the popular vote?
- What chances do the candidates have of winning some particular state?
- Who was winning on some particular date in the middle of the election?
- If the election were held today, who would win?
- Which states are most important for winning the election?
- How did particular major events affect the race?

In this document, we describe our work at a detailed non-technical level. We have tried to keep the prerequisite knowledge very low, perhaps knowledge of what a standard deviation is, but only the reader can judge the success of this effort. Should there be interest, we will write a more technical description of how we implement these ideas using the Kalman Filter to form estimates and make predictions. Until then, see [And], [Koo], and [Wik] for more information.

2 Guiding Principles

The main principle we follow is that our methodology should follow from data and sound mathematical theory to the greatest extent possible, and our results should follow directly from our methodology. Where we must make assumptions that do not flow from data and theory, we will make them explicit.

We will not alter results ad hoc; we have faith that the rigorous application of data and mathematics is more effective than our own hunches or even pundits'. When we discover errors on our method, we will prefer to fix those errors by correcting flaws in the underlying data, theory, or assumptions rather than adding hacks to our code.

While we openly support Senator Obama's candidacy, our results will be objective. We will scorn and mock any fool who announces that our results are biased, especially if they fail to suggest a flaw in our data, theory, or assumptions. On the other hand, we welcome suggestions for improvements in our assumptions or algorithm (although we won't have time to investigate or implement many of them).

3 Assumptions

This is a list of the fundamental assumptions that form the model we use. A longer list of more technical assumptions is in the appendix. We recognize that most of these assumptions are not quite correct, but believe that they are a close enough approximation to the truth that the model will work. Although the implications of some of these may not be obvious from the non-technical writeup, they do all have implications for the results, and we are happy to answer questions about how each of them gets used in our model.

- Popular opinion changes randomly over time. Therefore, all things being equal, a poll taken today is better than a poll taken a month ago, although we can glean information from both. Also, even a poll that describes today's opinion perfectly is an imperfect predictor of future opinion.
- In particular, popular opinion is not mean-reverting. That is, both candidates are equally likely to pick up support at any point in time, even if one is already doing better than members of his party usually do. (Note

that we don't actually believe that this assumption accurately reflects reality, but haven't found a good way to work mean-reversion into the model.)

- Changes to popular opinion in different states are strongly positively correlated.
- There is no momentum to changes in popular opinion. That is, if we discover that Obama gains 1 point one day, we still assume that he will be equally likely to gain or lose support the next day. Note: We believe this is false, and that introducing some serial correlation into the model would be a significant improvement. We would welcome suggestions about how to measure the serial correlation of changes in opinion, and how to introduce this new assumption into our calculations.
- Polls carry some inaccuracy because pollsters can only sample a very small number of people, typically in the range of 400–1000. This is called *sample error*.
- Polls also carry some inaccuracy from other sources. Some people may not be home when a pollster calls, others may be more or less willing to take a poll, others may not have a telephone, others may be confused about the questions, others may feel pressured to give a dishonest answer, etc. This is called *non-sample error*.
- Polls have equal amounts of non-sample error.
- Polls are unbiased.
- The likelihood of undecideds breaking one way or another is independent of how previously-decideds have already broken. (It is also independent of any poll's sample or non-sample error.) Note: Another reasonable assumption would be that undecideds tend to vote in the same proporations that previously-decided voters do. We would love to see evidence one way or the other on this.
- There is high uncertainty in how undecideds will break, and the allocation of undecideds in different states will be strongly positively correlated.

4 Creating and Combining Estimates

There's a lot of data out there. Some polls may say that Obama is ahead nationally, some may say that McCain is ahead nationally. Some may say that Obama has a landslide in Ohio, some may say that the race is a "statistical tie."[1] Some states may have many polls, some may have few. Some polls are taken nationally, some at the level of individual states. Some polls were taken a long time ago, some were taken recently. This section describes how we combine all this information into a coherent result.

4.1 Combining Multiple New Polls

The easiest case is how to combine similar polls, i.e., polls from the same population taken on the same day. The obvious solution is to add the responses of the polls together. Thus, if a 500-person poll reports that Obama has 50% of the vote and a 1500-person poll reports that Obama has 40% of the vote, then the two polls are equivalent to one 2000-person poll that reports that Obama has 42.5% of the vote. This is because a total of 850 people out of 2000 people said that they supported Obama.

We do something slightly more sophisticated than this in order to account for non-sample error. See below for details.

4.2 Combining a New Poll and an Existing Estimate

A typical poll says something of the form, "42.5% of a sample of 2000 people support Obama." A generic *a priori* estimate says something of the form, "We estimate that 45% of the population supports Obama, and our uncertainty is such that we consider a 4% error to be 1 standard deviation."[2]

[1] A dumb phrase, concocted no doubt by the media. When people try to define it, they usually make some noises about how statistical ties lack 95% certainty. In fact, even when they don't garble the phrasing completely, they always miss the facts that polls have non-sample error and that the uncertainty in the two candidates' numbers are very strongly negatively correlated. More fundamentally, there's a continuous spectrum between zero certainty and perfect certainty that "statistical tie"-centric reporting lacks

[2] That is to say, if we arrived at this same estimate many different times, we'd expect the actual results to be normally distributed with a mean of 45% and a standard deviation of 4%, so that we would be off by more than 4% about 68% of the time.

As it turns out, there's an easy and well-known way to translating between these statements. The poll in the first statement is equivalent to an estimate that says, "We estimate that 42.5% of the population supports Obama, and our uncertainty is such that we consider a 1.105% error to be 1 standard deviation." Here, the 1.105% arises as the square root of $(42.5\%)(100\% - 42.5\%)/2000$. Or, in reverse, the estimate in the second statement is equivalent to a poll that says, "45% of a sample of 154.7 people support Obama." Here, the 154.7 arises as $(45\%)(100\% - 45\%)$ divided by the square of the standard deviation.[3]

With both statements in "poll" format, it's easy to combine the two estimates. The two estimates add up to form an estimate equivalent to a single poll that says, "919.6 or 42.7% out of a sample of 2154.7 people support Obama."

In practice, the second "generic estimate" format will be much more useful for reading off results, and it will also be much more extensible when we need to think about correlations between states. So here's how to combine two estimates in the "generic estimate" format. If you have two estimates with expectations μ_1, μ_2 and standard deviations σ_1, σ_2, then your combined expectation is

$$\frac{\mu_1\sigma_2^2 + \mu_2\sigma_1^2}{\sigma_1^2 + \sigma_2^2},$$

and your combined standard deviation is

$$\frac{\sigma_1\sigma_2}{\sqrt{\sigma_1^2 + \sigma_2^2}}.$$

It's easy to check that these two ways of combining estimates give equivalent results.

4.3 Non-Sample Error

The naïve translation about takes only sample error. That is to say, if a 500-person poll has Obama getting 50% of the vote, then it predict a 2.236% standard deviation of sample error. As stated elsewhere, we believe that polls have other sources of error, and that non-sample error is uncorrelated to sample error. Therefore, the right way to combine sample and non-sample error is to add the variances.

[3]The square of the standard deviation is an important statistic known as the variance.

We don't have a good way of understanding how much non-sample error polls tend to have (although we hope to develop this), but suppose that polls carry a non-sample error with variance 0.0004. In our example, the sample error has variance 0.0005, so the total error has variance 0.0009, and the standard deviation of the poll's total error is 3%.

4.4 Interstate Inference

The basic insight here is simple: When public opinion in one state goes up, it's likely that public opinion in another state goes up as well. In the simplest case, suppose that we start out with estimates μ_1 and μ_2 in two states, and standard deviations σ_1, σ_2 on those estimates. Suppose also that the uncertainty in those states is 90% correlated. Now suppose a new poll (or maybe final election result) comes out that shows, with total certainty, that State 1 is one standard deviation higher than expectation. Because the two states are correlated, our estimate of State 2 should now have an expectation that is 0.9 standard deviations higher than it was before. (Our new estimate for State 2 also has a smaller standard deviation than our old estimate for State 2, which makes sense since we now have more information. I think in this case it would decrease by a factor of $\sqrt{0.9}$, but someone should check me on this.)

The great generalization of numbers into vectors and matrices provides a tool for implementing this logic (and to provide for the case where the new poll has uncertainty). Let's take another look at the equations we found for combining two estimates. Recall that we had an old estimate with expectation μ_o and standard deviation σ_o, and a new poll with expectation μ_p and standard deviation σ_p. Rearranging the terms a little, we get that the new combined expectation μ_n is given by

$$\mu_n - \mu_o = \frac{\sigma_o^2}{\sigma_o^2 + \sigma_p^2}(\mu_p - \mu_o),$$

and the new combined standard devation σ_n is given by

$$\sigma_n^2 - \sigma_o^2 = -\frac{\sigma_o^4}{\sigma_o^2 + \sigma_p^2}.$$

To see how this generalizes to the multi-state case, let us suppose instead that our old estimate is given by:

- a set of expectations for each state, presented as a column vector M with one entry for each state; and
- a matrix V to express the uncertainty we have about each state, and how those uncertainties are correlated. The diagonal entries are the variances of our estimates in each separate state, and the off-diagonal entries are the covariances between our estimates in different states. Technically, this means the (i, j)th entry is the correlation of the ith and jth states, times the geometric mean of the variances of the two states.

Note that M is analogous to the μ_o of the single-state example, and V is analogous to the σ_o^2 of the single-state example.

Suppose we have a poll in state i that suggests an expectation of μ_p with uncertainty σ_p. Let M_i be the ith entry of M, let V_i be the ith column of V, and let V_{ii} be the ith diagonal entry of V. Then the new estimate's expectation vector M_{new} is given by

$$M_{\text{new}} - M = V_i(V_{ii} + \sigma_p^2)^{-1}(\mu_p - M_i),$$

and the new estimate's variance matrix V_{new} is given by

$$V_{\text{new}} - V = -V_i(V_{ii} + \sigma_p^2)^{-1}V_i'.$$

(Here, V_i' represents the transpose of V_i, or the ith row of V.) It should be clear how this result directly generalizes the single-state formulas.

Another common case is when a poll is taken across a wide number of states, such as a national poll. (The following approach also applies to a poll in Maine or Nebraska, which span multiple congressional districts, each of which grants an electoral vote.) Let's take the example of a national poll. Let Z be a column vector that represents the proportion of the nation's voters in each state. Let the old estimate M, V be defined as before, and suppose that the national poll indicates a national expectation of μ_p with uncertainty σ_p. Then the new estimate's expectation vector M_{new} is given by

$$M_{\text{new}} - M = VZ(Z'VZ + \sigma_p^2)^{-1}(\mu_p - Z'M),$$

and the new estimate's uncertainty is given by

$$V_{\text{new}} - V = -VZ(Z'VZ + \sigma_p^2)^{-1}Z'V.$$

It should be clear again here how this result directly generalizes the previous result.

4.5 Aging Our Predictions

Since we assume that public opinion is non-mean-reverting and lacks momentum, our expectations do not change through time in the absence of new polling information. This explains why our chart of Obama's estimated lead through time becomes a flat line between today and election day. However, our uncertainty certainly does increase over time.

As stated elsewhere, we believe that public opinion changes every day, with constant volatility, and that changes on different days are uncorrelated. These assumptions imply that the variance of our uncertainty about Day 1 is equal to the variance of our uncertainty about Day 0, plus the variance of a one-day opinion change.

Of course, we assume something more about changes in public opinion, namely that states' changes are correlated. In other words, it's far more likely that McCain will improve in both OH and PA than that he will improve in one and falter in the other. This means that we have a whole variance-covariance matrix that represents a one-day opinion change. We add this matrix to the existing Day 0 uncertainty matrix to get an uncertaintly matrix for Day 1.

4.6 Undecideds

This is going to be filled in later. The main points are:

- We keep separate running track of the Obama numbers, McCain numbers, and undecided numbers.
- We assume that changes in these three numbers are negatively correlated at -50%, so that if the Obama numbers go up a little, then the other two numbers both likely go down.
- After running the whole information-combining process through to election day, allocate the remaining undecideds between Obama and McCain, increase the uncertainty for the Obama and McCain numbers (since there's uncertainty in how undecideds will ultimately break), and force the uncertainty of Obama + McCain to 0.

4.7 The Whole Process

We start with the results of the 2004 Bush–Kerry election at Day 0, and with 0 uncertainty for that day. Then for each day, we do the following loop:

1. Age the old estimate by one day by adding the variance of a one-day opinion change. This creates a new estimate with the same expectation and a slightly larger uncertainty.
2. Combine this estimate with a new poll to get a new estimate.
3. Repeat step 2 for each new poll.

Once we get to election day, we allocate undecideds as previously described.

The process described here, called Kalman filtering, outputs an estimate of public opinion on each day D based on the information available on or before day D. There's a similar process called Kalman smoothing that goes backward to output an estimate of public opinion on each day D based on all information available. It naturally incorporates the same assumptions about volatility and correlation as the filtering process does. This smoothing step isn't important for getting an estimate of what will happen on election day, but it's useful for understanding the story of the campaign.

The output of this is that, for each day of the campaign, we have a vector representing our estimate of each state's opinion, and a matrix representing our uncertainty about each state's opinion and the correlation between our uncertainties in each state.

5 Simulations

The process described in the last section outputs a vector that represents our best estimate of Obama's support on election day, and a matrix that represents our uncertainty about each state's opinion, together with the correlation between results in each states.

We generate a large number (usually 10000–30000) of sets of random variables with their distribution given by this vector and matrix. Each set of random variables represents the results of one simulation of the election. From this vector of results in each state, we can read off winner in each state, the electoral

college winner, and the popular vote result. From looking at the set of all simulations, we can read off the likelihood of victory, the likelihood of victory in each state, typical electoral vote totals, typical electoral college coalitions, and various measures of which states are important to the election results.

A Technical Assumptions

- Popular opinion is equally volatile at any point in time, and equally correlated between states at any point in time. This is probably false, but I don't have a good way to measure how volatility changes, and it's probably not a very important point. On the one hand, I've heard arguments that volatility increase near election day, since people start paying attention; on the other hand, I've heard arguments that volatility decreases near election day, since people already know everything they're going to know about the candidates.
- Volatility of popular opinion is equal in each state. (More precisely, in the 48 states besides ME and NE, in DC, and in each of the congressional districts of ME and NE.)
- Volatility in the percentage of voters going to Obama, McCain, and Undecided is equal, and the correlations between these three percentages are -50%. This -50% figure is what results from restricting a 3-variable standard normal distribution to the plane $X + Y + Z = 1$.
- National daily volatility is 0.000004. Note: This number is essentially an educated guess, and we have started to see evidence that this number is too low, and our next major project is to get a better, more principled estimate of this.
- The sample error of each poll contributes a variance of 0.0004. We assume that sample error and non-sample error are independent.
- The number of voters in 2008 will be the same as the number in 2004 who voter for Bush or Kerry. Note: We will soon change our code to reflect population changes since 2008; however, this change will have virtually no effect on popular vote calculations, and

- When pollsters do national (or other multi-jurisdiction polls), they sample states in proportion to the number of voters who will turn out in each state. (We don't know if this is actually true; it may be that pollsters sample in proportion to something else, like population or number of telephones. Let us know if you have some information about polling methods.)
- When a poll spans multiple days, we treat it is though it all happened on the middle day of the poll. When a poll spans an even number of days, we round forward in time.

B The Presidential Election Process

The process of electing the President is described by [Con] as below. For more information on the Electoral College, see [Kim].

> **Article II**
>
> **Section 1.** The executive Power shall be vested in a President of the United States of America. He shall hold his Office during the Term of four Years, and, together with the Vice President[4], chosen for the same term, be elected, as follows:
>
> Each State shall appoint, in such Manner as the Legislature thereof may direct[5], a Number of Electors, equal to the whole Number of Senators and Representatives to which the State may be entitled in the Congress; but no Senator or Representative, or Person holding an Office of Trust or Profit under the United States, shall be appointed an Elector....
>
> The Congress may determine the Time of chusing the Electors[6],

[4]The Vice President does two things: He is President of the Senate and breaks ties there; and he becomes President or acts as President when the President dies (becomes), resigns (becomes), is expelled from office (becomes) or becomes otherwise unable to serve (acts as). For the last possibility, see the 25th Amendment.

[5]Each state except Maine and Nebraska assigns all its electors to the winner of a statewide popular election. Maine and Nebraska both elect two such at-large electors, and one from each Congressional district.

[6]The first Tuesday after the first Monday of November in years divisible by four. [Kim] In 2008, this will be Nov. 4th

and the Day on which they shall give their Votes[7]; which Day shall be the same throughout the United States.

Amendment XII

The Electors shall meet in their respective states and vote by ballot for President and Vice President, one of whom, at least, shall not be an inhabitant of the same state with themselves; they shall name in their ballots the person voted for as President, and in distinct ballots the person voted for as Vice-President, and they shall make distinct lists of all persons voted for as President, and of all persons voted for as Vice-President, and of the number of votes for each, which lists they shall sign and certify, and transmit sealed to the seat of the government of the United States, directed to the President of the Senate[8];—The President of the Senate shall, in their presence of the Senate and House of Representatives, open all the certificates and the votes shall then be counted;[9]—The person having the greatest number of votes for President, shall be the President, if such number be a majority of the whole number of Electors appointed; and if no such number be a majority of the whole number of Electors appointed; and if no person have such majority, then from the persons having the highest numbers not exceeding three on the list of those voted for as President, the House of Representatives shall choose immediately, by ballot, the President. But in choosing the President, the votes shall be taken by states, the representation from each state having one vote; a quorum for this purpose shall consist of a member or members from two-thirds of the states, and a majority of all the states shall be necessary to a choice.... The person having the greatest number of votes as Vice-President, shall be the Vice-President, if such number be a majority of the whole number of Electors appointed, and if no person have a majority, then from the two highest numbers on the list, the Senate shall choose

[7]The first Monday after the second Wednesday in December following the popular election. [Kim] In 2008, this will be Dec. 15th.

[8]I.e., the Vice President of the United States.

[9]This happens on Jan. 6th after the election.

the Vice-President; a quorum for the purpose shall consist of two-thirds of the whole number of Senators, and a majority of the whole number shall be necessary to a choice. But no person constitutionally ineligible to the office of President shall be eligible to that of Vice-President of the United States.

Amendment XX

Section 3. If, at the time fixed for the beginning of the term of the President, the President elect shall have died, the Vice President elect shall become President. If a President shall not have been chosen before the time fixed for the beginning of his term, or if the President elect shall have failed to qualify, then the Vice President elect shall act as President until a President shall have qualified; and the Congress may by law provide for the case wherein neither a President elect nor a Vice President elect shall have qualified, declaring who shall then act as President, or the manner in which one who is to act shall be selected, and such person shall act accordingly until a President or Vice President shall have qualified.

Section 4. The Congress may by law provide for the case of the death of any of the persons from whom the House of Representatives may choose a President whenever the right of choice shall have devolved upon them, and for the case of the death of any of the persons from whom the Senate may choose a Vice President whenever the right of choice shall have devolved upon them.

Amendment XXII

Section 1. No person shall be elected to the office of the President more than twice, and no person who has held the office of President, or acted as President for more than two years of a term to which some other person was elected President shall be elected to the office of President more than once. . . .

Amendment XXIII

Section 1. The District constituting the seat of Government of the

United States[10] shall appoint in such manner as the Congress may direct:

A number of electors of President and Vice President equal to the whole number of Senators and Representatives in Congress to which the District would be entitled if it were a State, but in no event more than the least populous state; they shall be in addition to those appointed by the States, but they shall be considered, for the purposes of the election of President and Vice President, to be electors appointed by a state; and they shall meet in the District and perform such duties as provided by the twelfth article of amendment.

Amendment XXIV

Section 1. The right of citizens of the United States to vote in any primary or other election for President or Vice President, [or] for President and Vice President, . . . shall not be denied or abridged by the United States or any State by reason of failure to pay any poll tax or other tax.

Amendment XXVI

Section 1. The right of citizens of the United States, who are eighteen years of age or older, to vote shall not be denied or abridged by the United States or by any State on account of age.

C Disclaimer

This project contains solely the ideas, work, and opinions of Benjamin Schak except where explicitly noted. In particular, it has not been reviewed, endorsed, approved, or funded by any current or past employer.

References

[And] Anderson, Brian D. O. and John B. Moore. *Optimal Filtering*. Englewood Cliffs, N.J.: Prentice-Hall, 1979.

[10] I.e., Columbia.

[Con] The Constitution of the United States of America. 1787, as amended 1791–1992.

[Kim] Kimberling, William C. "The Electoral College." `http://www.fec.gov/pdf/eleccoll.pdf` (Jul. 2nd 2004).

[Koo] Koopman, Siem Jan. "Exact Initial Kalman Filtering and Smoothing for Nonstationary Time Series Models." *J. of the American Statistical Association*, Vol. 92, No. 440 (Dec. 1997), pp. 1630–38.

[Wik] Wikipedia. "Kalman filter." `http://en.wikipedia.org/wiki/Kalman_filter` (Jun. 22nd 2008).