Forecasting Elections with Non-Representative Polls

Wei Wang

Department of Statistics, Columbia University

July 6th, 2015

joint work with David Rothschild (MSR), Sharad (Stanford), and Andrew Gelman (Columbia).

Outline

- Non-Representative Polls in the Era of Big Data
- The Xbox Poll on 2012 Presidential Election
- Statistical Adjustments: Multilevel Regression and Poststratification
- Results
- Discussions

Big, but Non-representative

- Modern opinion polls are built on the premise of representative, probabilistic sampling.

Big, but Non-representative

- Modern opinion polls are built on the premise of representative, probabilistic sampling.
- Non-representative polls are generally considered unacceptable.

Big, but Non-representative

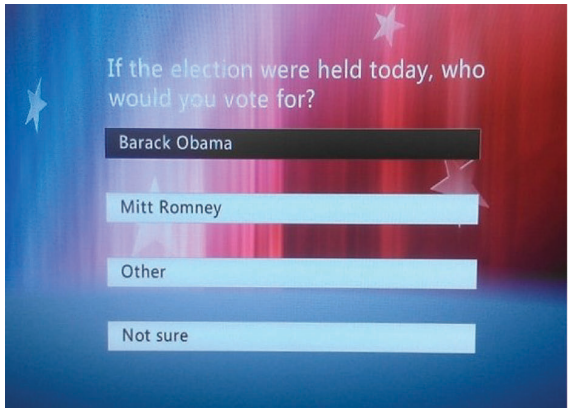
- Modern opinion polls are built on the premise of representative, probabilistic sampling.
- Non-representative polls are generally considered unacceptable.
- However, Big Data tend to be non-representative, convenient samples, i.e., huge selection bias.

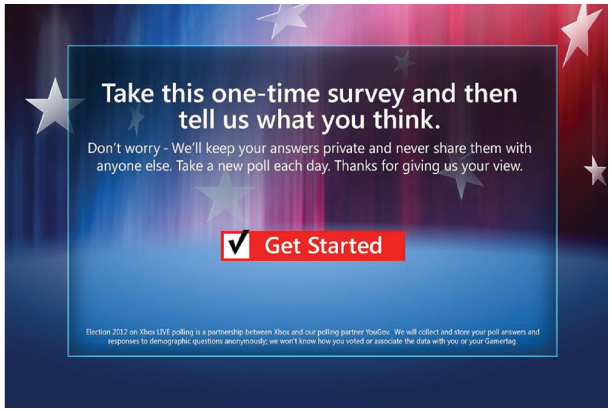
Big, but Non-representative

- Modern opinion polls are built on the premise of representative, probabilistic sampling.
- Non-representative polls are generally considered unacceptable.
- However, Big Data tend to be non-representative, convenient samples, i.e., huge selection bias.
- With massive amount of data, modern computing power and advanced statistics technology, non-representative polls might present a useful alternative for representative opinion polls.

Xbox Data

- Working with researchers from Microsoft, we placed an opt-in poll continuously available on the Xbox gaming platform during the 45 days preceding the 2012 U.S. presidential election.





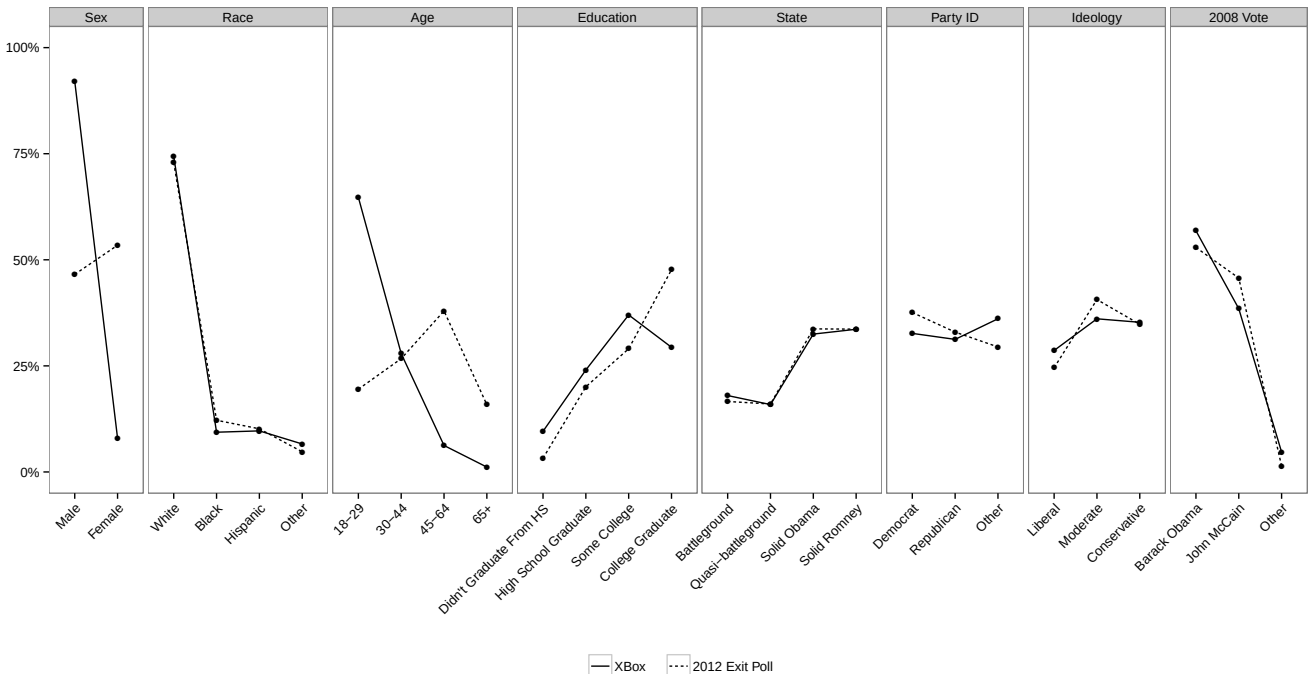
- There are 700k responses with 300k unique respondents.

Demographic Compositions

- Demographic information including sex, race, age, education, state, party ID, ideology and vote in the last election was collected once when respondents took the survey for the first time.

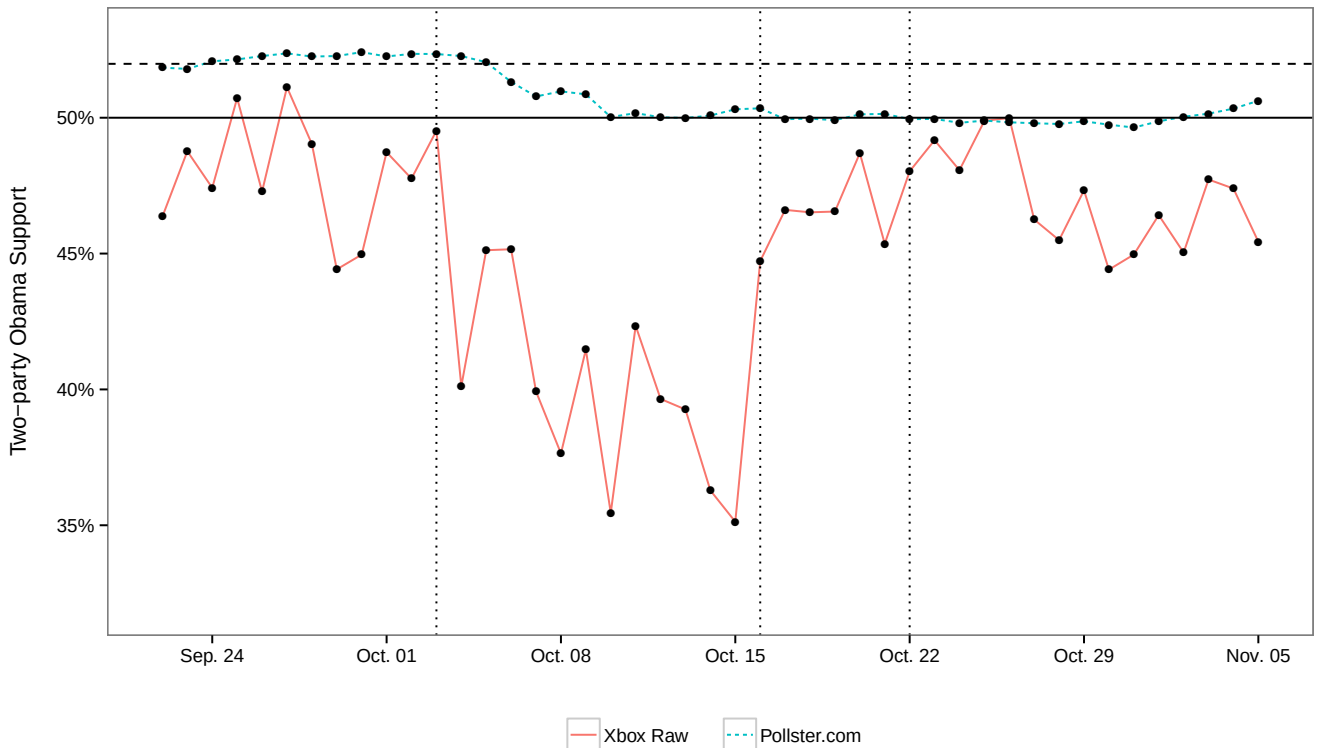
Demographic Compositions

- Demographic information including sex, race, age, education, state, party ID, ideology and vote in the last election was collected once when respondents took the survey for the first time.
- The demographic compositions differ greatly from the 2012 Exit Poll in gender and age.



Raw Results from Xbox

- The raw results suggest a Romney landslide.



Who is Mr.P?

- MRP stands for Multilevel Regression and Poststratification. It has been shown successful in gauging public opinions in political science literature (Buttice & Highton, 2013; Lax & Phlips 2009).
- Obviously, it is a two-step procedure, consisting of Multilevel Regression (Bayesian Hierarchical Models) and Poststratification.

Multilevel Regression: Goals

- Although a gamer is different from a voter, conditional on various demographic variables, the bias should be small, or even eliminated

$$P(\text{Gamer for Obama} \mid \text{demog}) \approx P(\text{Voter for Obama} \mid \text{demog})$$

Multilevel Regression: Goals

- Although a gamer is different from a voter, conditional on various demographic variables, the bias should be small, or even eliminated

  $$P(\text{Gamer for Obama} \mid \text{demog}) \approx P(\text{Voter for Obama} \mid \text{demog})$$

- The more demographic variables we use, the more reasonable the above claim, but also the smaller the cells defined by these demographic variables.
- It doesn't help that real-world data tend to be highly skewed, thus the sparsity.

Multilevel Regression: Goals

- Although a gamer is different from a voter, conditional on various demographic variables, the bias should be small, or even eliminated

  $$P(\text{Gamer for Obama} \mid \text{demog}) \approx P(\text{Voter for Obama} \mid \text{demog})$$

- The more demographic variables we use, the more reasonable the above claim, but also the smaller the cells defined by these demographic variables.
- It doesn't help that real-world data tend to be highly skewed, thus the sparsity.
- How can we get stable estimates at sparse cells?

Multilevel Regression: Models

- Multilevel Regression, or Hierarchical Models, mitigates the overfitting problem by assigning priors on group-level parameters. In frequentist terms, this amounts to imposing regularizations.

Multilevel Regression: Models

- Multilevel Regression, or Hierarchical Models, mitigates the overfitting problem by assigning priors on group-level parameters. In frequentist terms, this amounts to imposing regularizations.
- We use multilevel models (hierarchical models) to estimate the cell-level responses.

$$\Pr(Y_i = \text{Obama}\}) = \text{logit}^{-1}\big(\beta_0 + \beta_1(\text{state last vote share}) + b^{\text{state}}_{j[i]} + b^{\text{edu}}_{j[i]} + b^{\text{sex}}_{j[i]} + b^{\text{age}}_{j[i]} + b^{\text{race}}_{j[i]} + b^{\text{party ID}}_{j[i]} + b^{\text{ideology}}_{j[i]} + b^{\text{last vote}}_{j[i]}\big)$$

with priors

$$b^{\text{var}}_{j[i]} \sim N(0, \eta^2_{\text{var}}),$$
$$\eta^2_{\text{var}} \sim \text{inv-}\chi^2(\mu, \eta^2_0).$$

Poststratification

- Poststratification is a common technique in survey sampling. It reweights subgroup-level estimates to obtain higher level estimates, e.g., state and national level estimates.

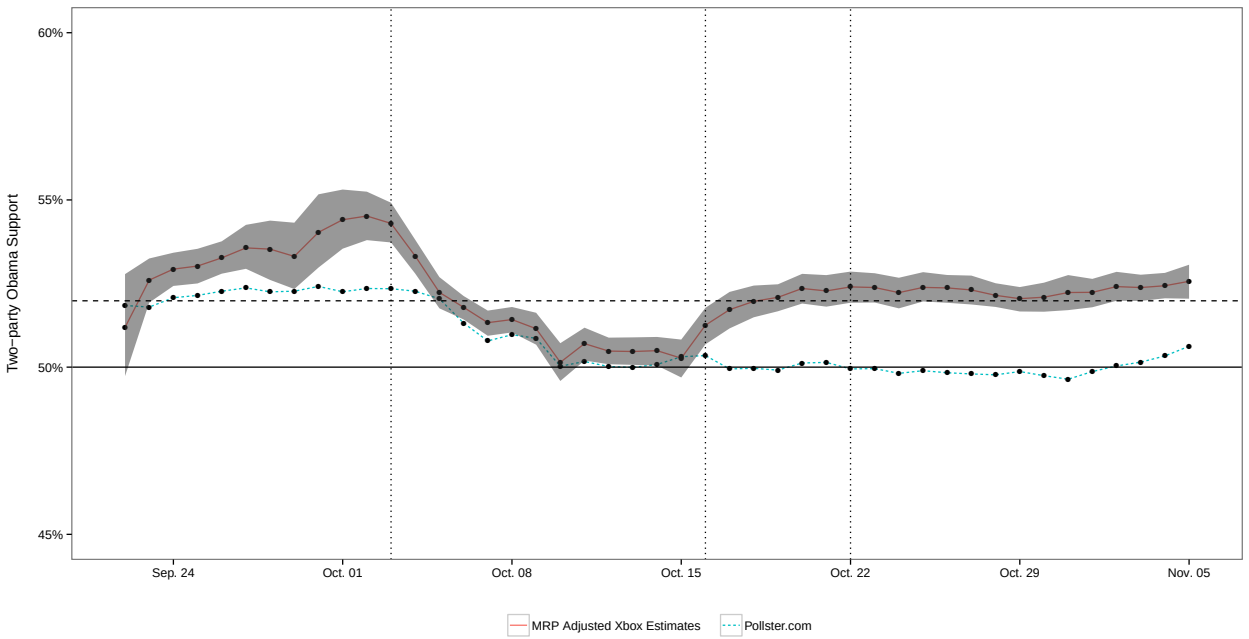
Poststratification

- Poststratification is a common technique in survey sampling. It reweights subgroup-level estimates to obtain higher level estimates, e.g., state and national level estimates.
- Once the estimates within each cell from Multilevel Regression step are in order, we can then the cell-level estimates up to a population-level by poststratification (in reference to 2008 Exit Poll data).

$$\hat{y}_{\text{PS}} = \frac{\sum_{j=1}^{J} N_j \hat{y}_j}{\sum_{j=1}^{J} N_j}$$

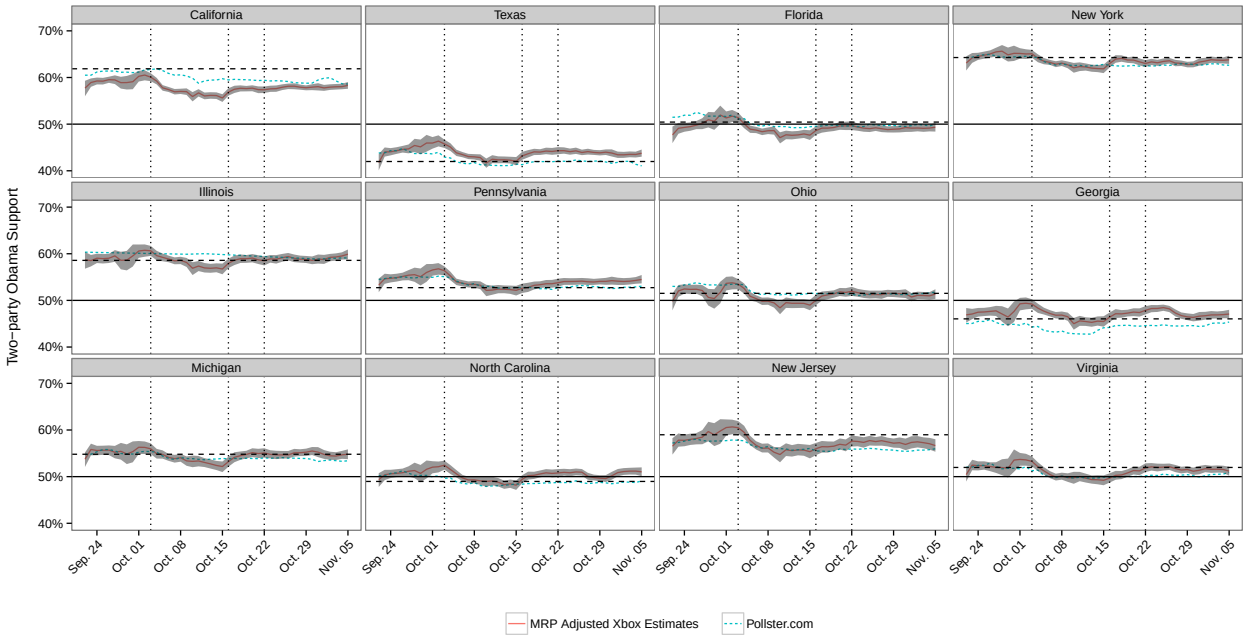
National Daily Snapshots

- The MRP adjusted daily snapshots provides a much reasonable time line of Obama two-party support during the 45 days period.



State Races Daily Snapshots

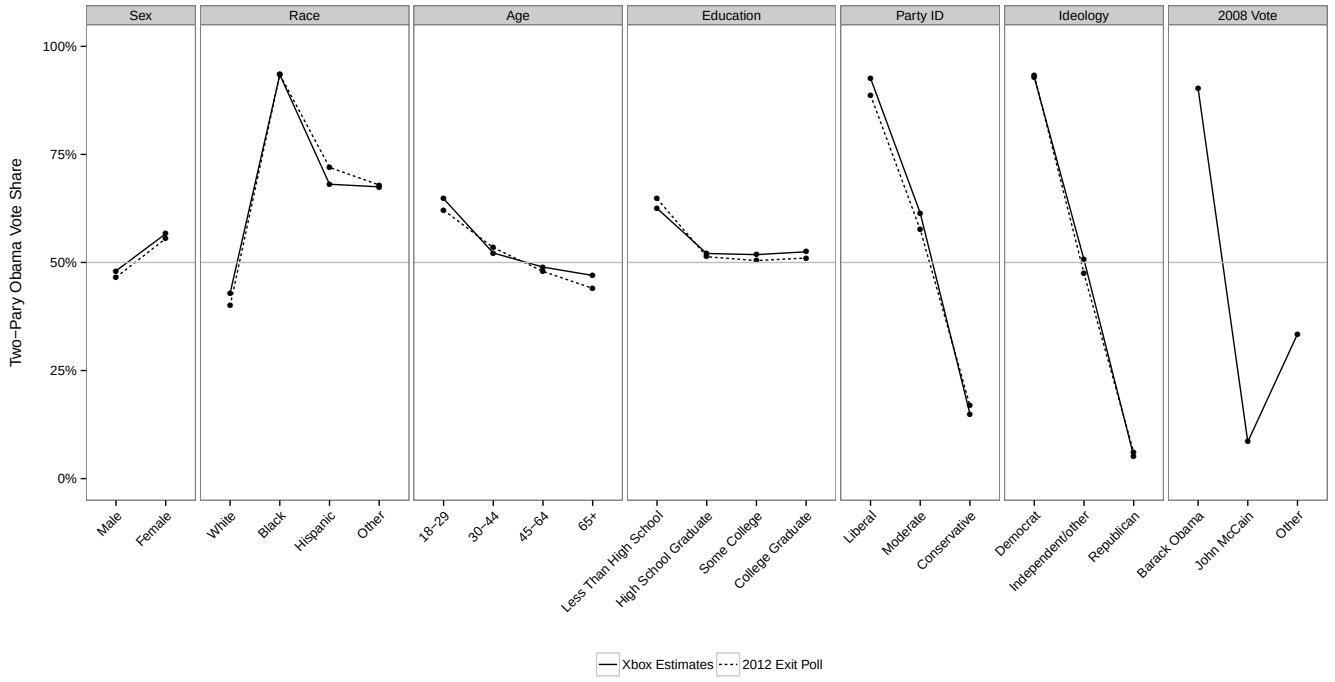
- National vote share is only of marginal importance in presidential election; to predict the election, we need to look at individual state races.



- The mean and median absolute errors of our estimates across 51 races on the day before the election are just 2.5 and 1.8 percentage points, respectively.

Demographic Subgroups

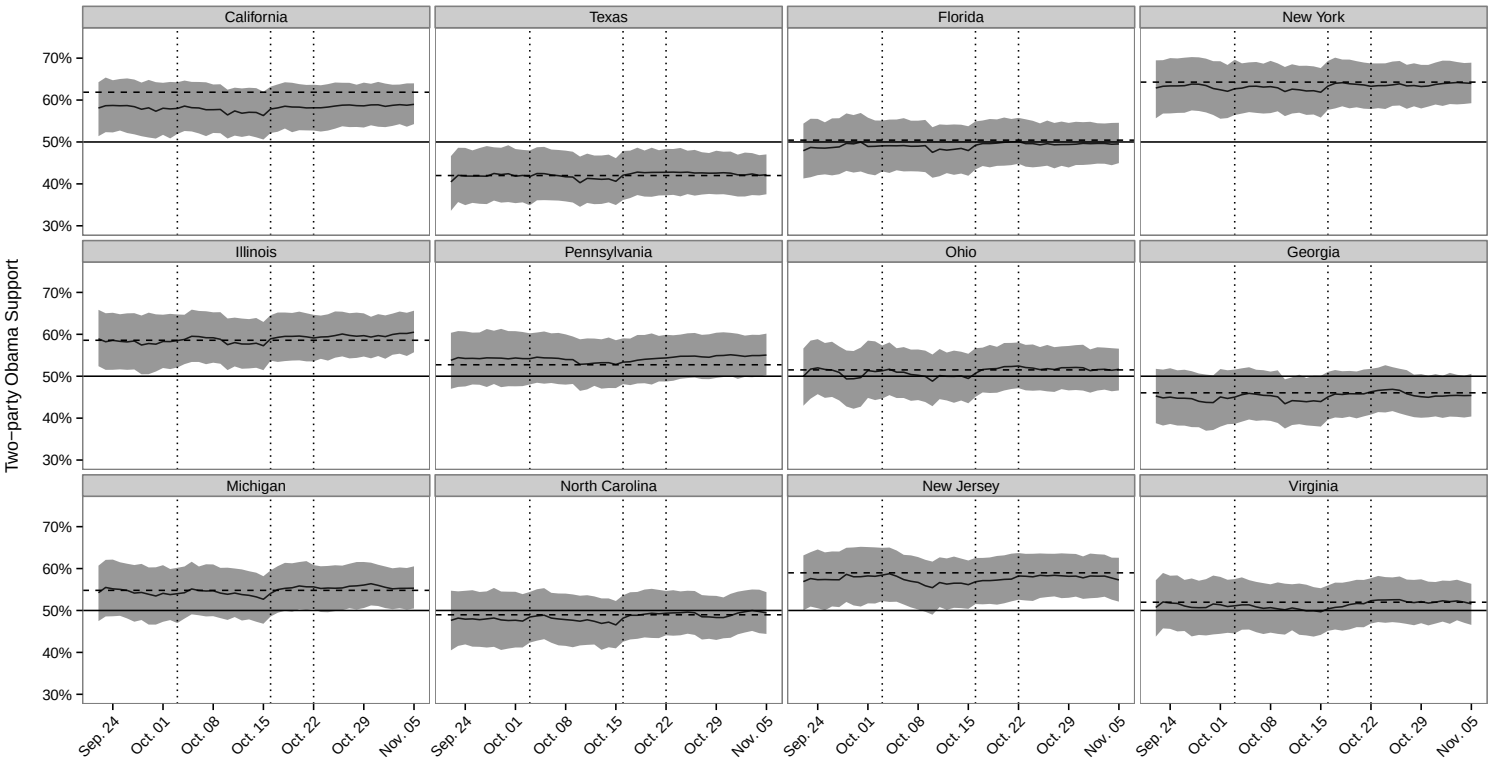
- We further look at the last day snapshot breaking down by demographic subgroups. The Xbox estimates are remarkably accurate, with a median absolute difference of 1.5 percentage points.



Election Day Outcome Calibrations

- Daily estimates of voter intent don't translate exactly to election day vote share estimates; some calibrations are needed.
- We collect historical daily topline polling results from 2000, 2004 and 2008 election, and run a regression model with time and daily voter's intent, and then apply the fitted model on our daily voter intent estimates to give election day vote share estimates.

State Race Results after Calibration



Electoral College Votes

- Since we have posterior distribution of the results of the electoral college races, we can find the posterior distribution of Obama's electoral college votes.

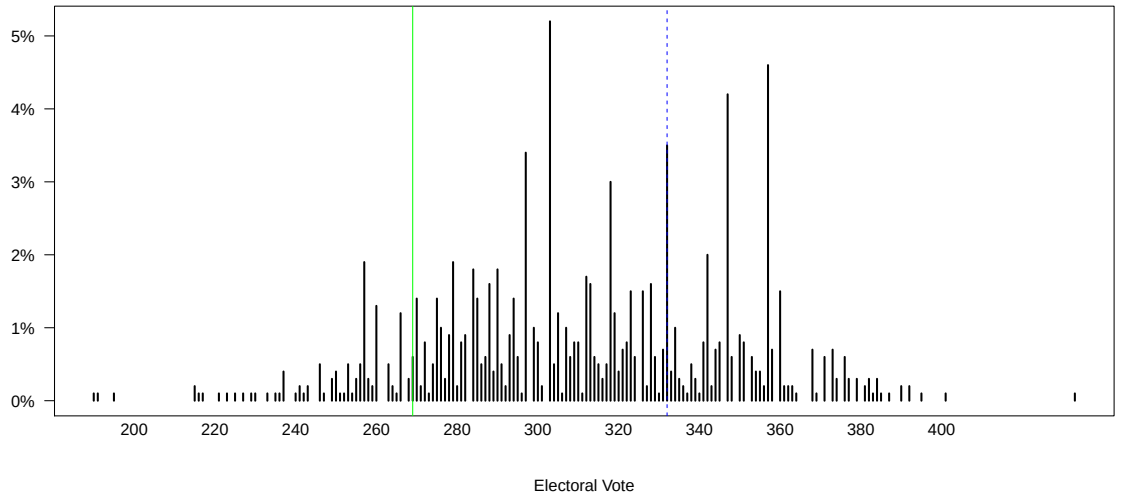


Figure : The green line represents 269, the minimum number that Obama needs for a tie. The blue line gives 332, the actual number of electoral votes captured by Obama. The estimated likelihood of Obama winning the electoral vote is 88%.

Discussions

- Representative sampling is still important and relevant in measuring public opinions. But with the advent of big data, low cost non-representative sampling could be an useful alternative.

Discussions

- Representative sampling is still important and relevant in measuring public opinions. But with the advent of big data, low cost non-representative sampling could be an useful alternative.
- But the use of non-representative sampling should be strictly guided by appropriate statistical methods, such as hierarchical modeling.

Discussions

- Representative sampling is still important and relevant in measuring public opinions. But with the advent of big data, low cost non-representative sampling could be an useful alternative.
- But the use of non-representative sampling should be strictly guided by appropriate statistical methods, such as hierarchical modeling.
- Non-representative sampling will be quite attractive in the cases where representative sampling might be too expensive or suffer from excruciatingly high non-response rate.