Disentangling Bias and Variance in Election Polls

Houshmand Shirani-Mehr
Stanford University

David Rothschild
Microsoft Research

Sharad Goel
Stanford University

Andrew Gelman
Columbia University

4 Jul 2017

Abstract

It is well known among researchers and practitioners that election polls suffer from a variety of sampling and nonsampling errors, often collectively referred to as *total survey error*. Reported margins of error typically only capture sampling variability, and in particular, generally ignore nonsampling errors in defining the target population (e.g., errors due to uncertainty in who will vote). Here we empirically analyze 4,221 polls for 608 state-level presidential, senatorial, and gubernatorial elections between 1998 and 2014, all of which were conducted during the final three weeks of the campaigns. Comparing to the actual election outcomes, we find that average survey error as measured by root mean square error (RMSE) is approximately 3.5 percentage points, about twice as large as that implied by most reported margins of error. We decompose this survey error into bias and variance components. We find average absolute election-level bias is about 1.5 percentage points, indicating that polls for a given election often share a common component of error. This shared error may stem from the fact that polling organizations often face similar difficulties in reaching various subgroups of the population, and rely on similar screening rules when estimating who will vote. We conclude by discussing how these results help explain polling failures in the 2016 U.S. presidential election, and offer recommendations to improve polling practice.

1 Introduction

Election polling is arguably the most visible manifestation of statistics in everyday life, and embodies one of the great success stories of statistics: random sampling. As is recounted in so many textbooks, the huge but uncontrolled Literary Digest poll of 1936 was trounced by Gallup's small, nimble quota sample, opening the door to modern sampling methods in election forecasting. Election polls are a high-profile reality check on statistical methods.

It has long been known that the margins of errors provided by survey organizations, and reported in the news, understate the total survey error. This is an important topic in sampling but is difficult to address in general for two reasons. First, we like to decompose error into bias and variance, but this can only be done with any precision if we have a large number of surveys and outcomes (not merely a large number of respondents in an individual survey). Second, assessment of error requires a ground truth for comparison, which is typically not available, as the reason for conducting a sample survey in the first place is to estimate some population characteristic that is not already known.

In the present paper we decompose survey error in a large set of state-level pre-election polls. This dataset resolves both of the problems just noted. First, the combination of multiple elections and many states gives us a large sample of polls. Second, we can compare the polls to actual election results.

1.1 Background

Modern election polls typically survey a random sample of eligible or likely voters, and then generate population-level estimates by taking a weighted average of responses, where the weights are designed to correct for known differences between sample and population.[1] This general analysis framework yields not only a point estimate of the election outcome, but also an estimate of the error in that prediction due to sample variance which accounts for

[1] One common technique for setting survey weights is raking, in which weights are defined so that the weighted distributions of various demographic features (e.g., age, sex, and race) of respondents in the sample agree with the marginal distributions in the target population [Voss, Gelman, and King, 1995].

the survey weights [Lohr, 2009]. In practice, weights in a sample tend to be approximately equal, and so most major polling organizations simply report 95% margins of error identical to those from simple random sampling (SRS) without incorporating the effect of the weights, for example ±3.5 percentage points for an election survey with 800 people.[2]

Though this approach to quantifying polling error is popular and convenient, it is well known by both researchers and practitioners that discrepancies between poll results and election outcomes are only partially attributable to sample variance [Ansolabehere and Belin, 1993]. As observed in the extensive literature on *total survey error* [Biemer, 2010, Groves and Lyberg, 2010], there are at least four additional types of error that are not reflected in the usually reported margins of error: frame, nonresponse, measurement, and specification. Frame error occurs when there is a mismatch between the sampling frame and the target population. For example, for phone-based surveys, people without phones would never be included in any sample. Of particular import for election surveys, the sampling frame includes many adults who are not likely to vote, which pollsters recognize and attempt to correct for using likely voters screens, typically estimated with error from survey questions. Nonresponse error occurs when missing values are systematically related to the response. For example, supporters of the trailing candidate may be less likely to respond to surveys [Gelman, Goel, Rivers, and Rothschild, 2016]. With nonresponse rates exceeding 90% for election surveys, this is a growing concern [Pew Research Center, 2016]. Measurement error arises when the survey instrument itself affects the response, for example due to order effects [McFarland, 1981] or question wording [Smith, 1987]. Finally, specification error occurs when a respondent's interpretation of a question differs from what the surveyor intends to convey (e.g., due to language barriers). In addition to these four types of error

[2] For the 19 ABC, CBS, and Gallup surveys conducted during the 2012 election and deposited into Roper Center's iPoll, when weights in each survey were rescaled to have mean 1, the median respondent weight was 0.73, with an interquartile range of 0.45 to 1.28. For a sampling of 96 polls for 2012 Senate elections, only 19 reported margins of error higher than what one would compute using the SRS formula, and 14 of these exceptions were accounted for by YouGov, an internet poll that explicitly inflates variance to adjust for the sampling weights. Similarly, for a sampling of 36 state-level polls for the 2012 presidential election, only 9 reported higher-than-SRS margins of error.

common to nearly all surveys, election polls suffer from an additional complication: shifting attitudes. Whereas surveys typically seek to gauge what respondents will do on election day, they can only directly measure current beliefs.

In contrast to errors due to sample variance, it is difficult—and perhaps impossible—to build a useful and general statistical theory for the remaining components of total survey error. Moreover, even empirically measuring total survey error can be difficult, as it involves comparing the results of repeated surveys to a ground truth obtained, for example, via a census. For these reasons, it is not surprising that many survey organizations continue to use estimates of error based on theoretical sampling variation, simply acknowledging the limitations of the approach. Indeed, Gallup [2007] explicitly states that their methodology assumes "other sources of error, such as nonresponse, by some members of the targeted sample are equal," and further notes that "other errors that can affect survey validity include measurement error associated with the questionnaire, such as translation issues and coverage error, where a part or parts of the target population ... have a zero probability of being selected for the survey."

1.2 Our study

Here we empirically and systematically study error in election polling, taking advantage of the fact that multiple polls are typically conducted for each election, and that the election outcome can be taken to be the ground truth. We investigate 4,221 polls for 608 state-level presidential, senatorial, and gubernatorial elections between 1998 and 2014, all of which were conducted in the final three weeks of the election campaigns. By focusing on the final weeks of the campaigns, we seek to minimize the impact of errors due to changing attitudes in the electorate, and hence to isolate the effects of the remaining components of survey error.

We find that the average difference between poll results and election outcomes—as measured by RMSE—is 3.5 percentage points, about twice the error implied by most reported confidence intervals.[3] To decompose this survey error into election-level bias and variance terms, we apply hierarchical Bayesian latent variable models [Gelman and Hill, 2007]. We find that average absolute election-level bias is about 1.5 percentage points, indicating that polls for a given election often share a common component of error. This result is likely driven in part by the fact that most polls, even when conducted by different polling organizations, rely on similar likely voter models, and thus surprises in election day turnout can have comparable effects on all the polls. Moreover, these correlated frame errors extend to the various elections—presidential, senatorial, and gubernatorial—within a state.

2 Data description

Our primary analysis is based on 4,221 polls completed during the final three weeks of 608 state-level presidential, senatorial, and gubernatorial elections between 1998 and 2014. Polls are typically conducted over the course of several days, and following convention, we throughout associate the "date" of the poll with the last day in which it was in the field. We do not include House elections in our analysis since polling is only available for a small and nonrepresentative subset of such races.

To construct the dataset, we started with the 4,154 state-level polls for elections in 1998–2013 that were collected and made available by FiveThirtyEight, all of which were completed during the final three weeks of the campaigns. We augmented these polls with the 67 corresponding ones for 2014 posted on Pollster.com, where for consistency with the FiveThirtyEight data, we considered only those completed in the last three weeks of the campaigns. In total, we ended up with 1,646 polls for 241 senatorial elections, 1,496 polls for 179 state-level presidential elections, and 1,079 polls for 188 gubernatorial elections.

In addition to our main dataset described above, we also consider 7,040 polls completed

[3] Most reported margins of error assume estimates are unbiased, and report 95% confidence intervals of approximately ±3.5 percentage points for a sample of 800 respondents. This in turn implies the RMSE for such a sample is approximately 1.8 percentage points, approximately half of our empirical estimate of RMSE.

during the last 100 days of 314 state-level presidential, senatorial, and gubernatorial elections between 2004 and 2012. All polls for this secondary dataset were obtained from Pollster.com and RealClearPolitics.com. Whereas this complementary set of polls covers only the more recent elections, it has the advantage of containing polls conducted earlier in the campaign cycle.

3 Summary statistics

For each poll in our primary dataset (polls conducted during the final three weeks of the campaign), we estimate total survey error by computing the difference between: (1) support for the Republican candidate in the poll; and (2) the final vote share for that candidate on election day. As is standard in the literature, we consider *two-party* poll and vote share: we divide support for the Republican candidate by total support for the Republican and Democratic candidates, excluding undecideds and supporters of any third-party candidates.

Figure 1 shows the distribution of these differences, where positive values on the x-axis indicate the Republican candidate received more support in the poll than in the election. For comparison, the dotted lines shows the theoretical distribution of polling errors assuming simple random sampling (SRS). Specifically, for each poll $i = 1, \ldots, N$, we simulate a polling result by drawing a sample from a binomial distribution with parameters n_i and $v_{r[i]}$, where n_i is the number of respondents in poll i who express a preference for one of the two major-party candidates, and $v_{r[i]}$ is the final two-party vote share of the Republican candidate in the corresponding election. We repeat this process separately for senatorial, gubernatorial, and presidential polls.

The plot highlights two points. First, for all three political offices, polling errors are approximately centered at zero. Thus, at least across all the elections and years that we consider, polls are not systematically biased toward either party. Indeed, it would be surprising if we had found systematic error, since pollsters are highly motivated to notice and correct

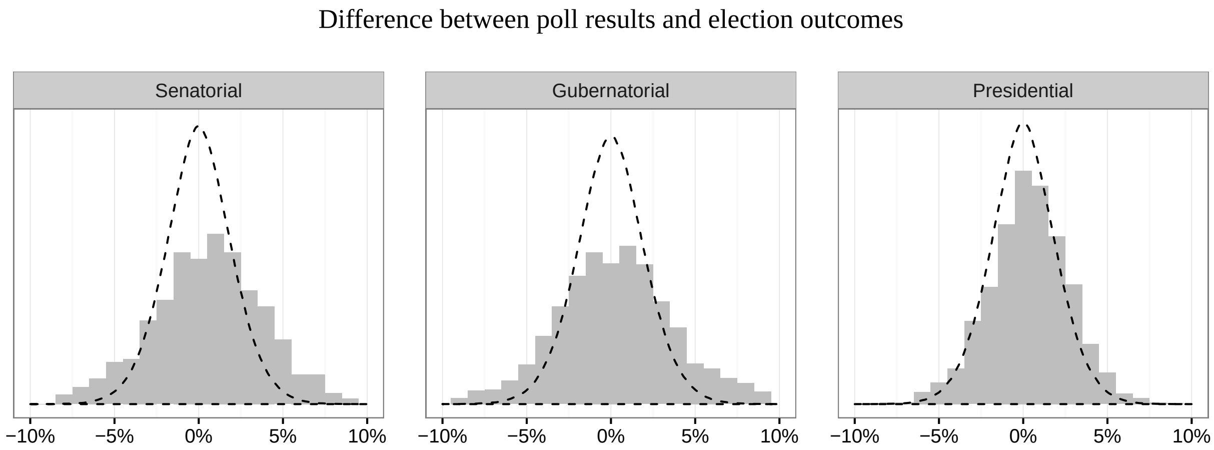

Figure 1: *The distribution of polling errors (Republican share of two-party support in the poll, minus Republican share of the two-party vote in the election) for state-level presidential, senatorial, and gubernatorial election polls between 1998 and 2014. Positive values indicate the Republican candidate received more support in the poll than in the election. For comparison, the dashed lines shows the theoretical distribution of polling errors assuming each poll is generated via simple random sampling. The polls in aggregate show no consistent bias toward either party, but they show a wider range of errors that would be expected from sampling variation alone.*

for any such aggregate bias. Second, the polls exhibit substantially larger errors than one would expect from simple random sampling. For example, it is not uncommon for senatorial and gubernatorial polls to miss the election outcome by more than 5 percentage points, an event that would rarely occur if respondents were simple random draws from the electorate.

We quantify these polling errors in terms of the root mean square error (RMSE).[4] The senatorial and gubernatorial polls, in particular, have substantially larger RMSE (3.7% and 3.9%, respectively) than SRS (1.9%). In contrast, the RMSE for state-level presidential polls is 2.5%, only a small amount higher than what one would expect from SRS. Because reported margins of error are typically derived from theoretical SRS error rates, the traditional intervals are too narrow. Namely, SRS-based 95% confidence intervals cover the actual outcome for only 73% of senatorial polls, 74% of gubernatorial polls, and 88% of presidential polls. It

[4] For each poll $i \in \{1, \ldots, N\}$, let y_i denote the two-party support for the Republican candidate, and let $v_{r[i]}$ denote the final two-party vote share of the Republican candidate in the corresponding election $r[i]$. Then RMSE is $\sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - v_{r[i]})^2}$.

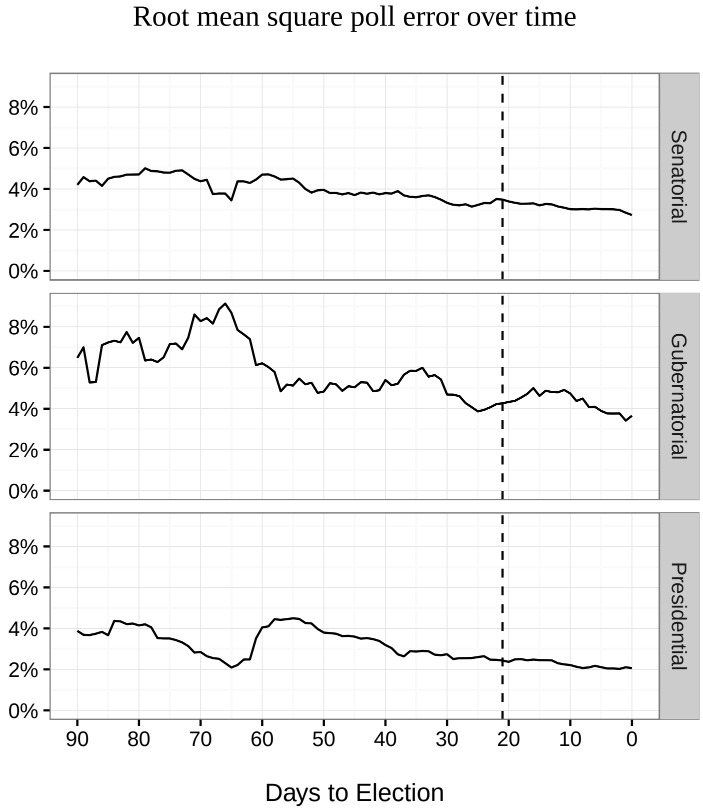

Figure 2: *Poll error, as measured by RMSE, over the course of elections. The graph for each day x indicates the root mean squared error for polls completed in the seven-day window centered at x. The dashed vertical line at the three-week mark shows that poll error is relatively stable during the final stretches of the campaigns, suggesting that the discrepancies we see between poll results and election outcomes are by and large not due to shifting attitudes in the electorate.*

is not immediately clear why presidential polls fare better; one possibility is that preferences in the presidential election are more stable, thus the survey response is a better predictor of vote choice; another is that turnout in the presidential election is easier to predict and so these polls suffer less from frame error.

We have thus far focused on polls conducted in the three weeks prior to election day, in an attempt to minimize the effects of error due to changing attitudes in the electorate. To examine the robustness of this assumption, we now turn to our secondary polling dataset and, in Figure 2, plot average poll error as a function of the number of days to the election. Due to the relatively small number of polls conducted on any given day, we include in each point in the plot all the polls completed in a seven-day window centered at the focal date (i.e., polls completed within three days before or after that day). As expected, polls early in the campaign season indeed exhibit more error than those taken near election day. Average error, however, appears to stabilize in the final weeks, with little difference in RMSE one month before the election versus one week before the election. Thus, the polling errors that we see during the final weeks of the campaigns are likely not driven by changing attitudes, but rather result from a combination of frame and nonresponse error. Measurement and specification error also likely play a role, though election polls are arguably less susceptible to such forms of error.

In principle, Figure 1 is consistent with two distinct possibilities. On one hand, election polls may typically be unbiased but have large variance; on the other hand, polls may have nonzero biases, but in aggregate these biases cancel to yield the depicted distribution. To determine which of these alternatives is driving our results, we decompose the observed poll error into election-level bias and variance components. The bias term captures systematic errors shared by all polls in the election (e.g., due to shared frame errors), while the variance term captures traditional sampling variation as well as variation due to differing survey methodologies across polls and polling organizations.

We start by assuming that poll results in each election r are independent draws from an unknown, election-specific *poll distribution* with mean μ_r and standard deviation σ_r. This poll distribution reflects both the usual sampling variation, as well as uncertainty arising from nonresponse, frame, and other sources of polling error. Our first goal is to estimate average absolute poll bias across the races in our dataset, where we separately consider senatorial, gubernatorial, and presidential elections. Specifically, we seek to estimate,

$$\frac{1}{k}\sum_{r=1}^{k}|b_r| = \frac{1}{k}\sum_{r=1}^{k}|\mu_r - v_r|,$$

where $|b_r| = |\mu_r - v_r|$ is the absolute bias in election r, and v_r is the final two-party vote

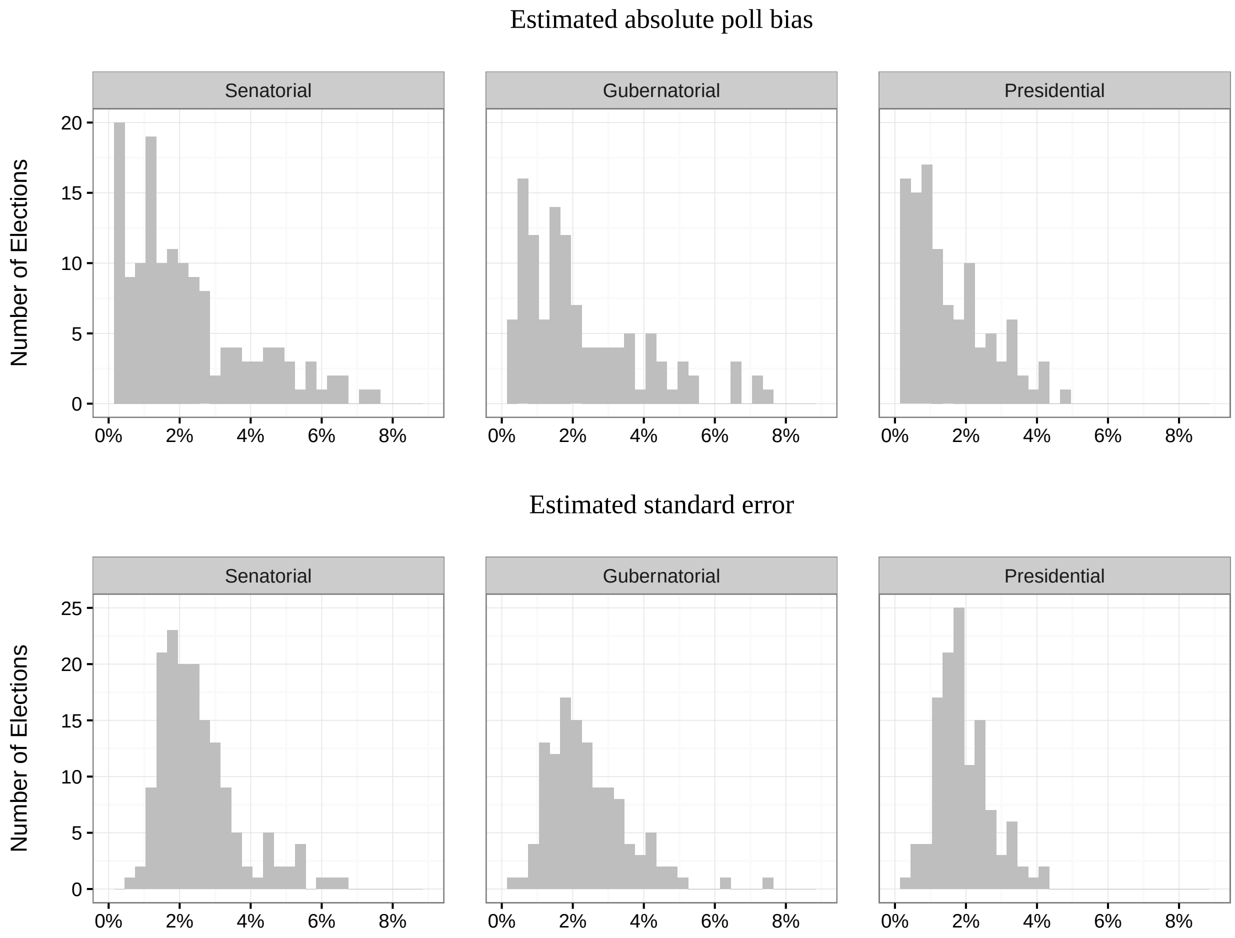

Figure 3: *Simple estimates of election-level absolute bias and standard error, obtained by taking the absolute difference between the average of polls in an election and the election outcome (top plot), and the sample standard deviation of polls in an election (bottom plot).*

share of the Republican candidate in that election. Our second goal is to estimate average standard error: $(1/k)\sum_{r=1}^{k}\sigma_r$.

A simple and intuitive estimate of election-level absolute poll bias $|b_r|$ is the absolute difference between the average of the poll results in that election and the election outcome itself. Similarly, we can approximate the standard deviation σ_r of the election-specific poll distribution by the sample standard deviation of the observed poll results. Specifically, suppose S_r is the set of polls that are conducted in election r. For each poll $i \in S_r$, denote by y_i the two-party Republican support in the poll, and denote by n_i the number of

	Senatorial	Gubernatorial	Presidential
Average error (RMSE)	3.7%	3.9%	2.5%
Average election-level absolute bias	2.1%	2.3%	1.4%
Average election-level standard error	2.5%	2.4%	1.9%

Table 1: *Simple estimates of RMSE, election-level bias, and election-level variance. In particular, election-level bias is estimated by taking the difference between the average of polls in a election and the election outcome. For reference, if polls were generated via simple random sampling, we would have average RMSE of 1.9%, zero average absolute bias, and 1.9% average standard error.*

respondents in the poll who express a preference for one of the two major-party candidates. Then a simple estimate of absolute poll bias is,

$$|\hat{b}_r| = \left| v_r - \frac{1}{|S_r|} \sum_{i \in S_r} y_i \right|, \tag{1}$$

and a simple estimate of variance is

$$\hat{\sigma}_r^2 = \frac{1}{|S_r| - 1} \sum_{i \in S_r} \left(y_i - \frac{1}{|S_r|} \sum_{i \in S_r} y_i \right)^2. \tag{2}$$

Returning to our primary dataset of polls completed within the final three weeks of the campaigns, we compute election-level absolute bias and variance for the 397 races for which we have at least four polls. Figure 3 shows the resulting distribution of estimates across races. Poll bias—particularly for senatorial and gubernatorial races—is often substantial, at times in excess of 5%. The election-level standard error of polls is likewise larger than what one would expect if polls were generated via SRS. As summarized in Table 1, this approach yields estimates of average absolute bias in senatorial and gubernatorial races of more than 2 percentage points, and average absolute bias of 1.4 percentage points in presidential races. The poll bias, which is not reflected in traditional margins of error, is estimated to be as big as the theoretical sampling variation from SRS.

4 Multilevel model for sampling and non-sampling variation

Analyzing a series of opinion polls can be seen as a form of meta-analysis, in which there is a single overall target which is estimated many times, with each estimate having its own bias, and with the distribution of these biases estimated from data. The standard approach to estimating these components of variation is multilevel modeling, and that is what we shall do here. For each poll i in election $r[i]$, let y_i denote the two-party support for the Republican candidate (as measured by the poll), where the poll has n_i respondents with preference for one of the two major candidates. Let $v_{r[i]}$ denote the final two-party vote share for the Republican candidate. We model the poll outcome y_i as a random draw from a binomial distribution parameterized as follows:

$$\begin{aligned} y_i &\sim \text{Binomial}\,(p_i\,,\, n_i) \\ \text{logit}(p_i) &= \text{logit}(v_{r[i]}) + \alpha_{r[i]} + \gamma_i \\ \alpha_j &\sim \text{N}(\mu_\alpha\,,\, {\sigma_\alpha}^2) \\ \gamma_i &\sim \text{N}(0\,,\, {\sigma_\gamma}^2) \end{aligned}$$

where $\alpha_{r[i]}$ and γ_i are the election-level and poll-level bias components on the logit scale.[5] Weakly informative priors are assigned to the hyper-paramaters μ_α, σ_α and σ_γ. Specifically, we set $\mu_\alpha \sim \text{N}(0, 0.2^2)$, $\sigma_\alpha \sim \text{N}_+(0, 0.2^2)$ and $\sigma_\gamma \sim \text{N}_+(0, 0.2^2)$. We fit this model separately for senatorial, presidential and gubernatorial elections.

Our priors are weakly informative in that they allow for a large, but not extreme, range of parameter values. In particular, 0.2 on the logit scale corresponds to approximately 0.05 on the probability scale, and although a 5 percentage point poll bias would be substantial, it is of approximately the right order of magnitude. The inverse gamma distribution is a traditional

[5] Given the relative stability of late-season polls (Figure 2), for simplicity we exclude a time component from this model. Our results are similar if we add an extra bias term for time.

choice of prior for variance parameters, but it rules out values near zero [Gelman, 2006]; our use of half-normal distributions is thus more consistent with our decision to select weakly informative priors. In Section 4 below, we experiment with alternative prior structures and show that our results are robust to the exact specification. Posterior distributions for the parameters are obtained via Hamiltonian Monte Carlo [Hoffman and Gelman, 2014] as implemented in Stan, an open-source modeling language for full Bayesian statistical inference [Stan Development Team, 2017].

The fitted model lets us estimate two key quantities. First, we estimate the average absolute poll bias μ_p by:

$$\hat{\mu}_p = \frac{1}{n}\sum_{i=1}^{n} |\hat{p}_i - v_{r[i]}|,$$

where $\hat{p}_i$ is the posterior mean of the p_i, the Republican share of the two-party vote in the population estimated from survey i. Second, we estimate the average absolute election bias μ_e by:

$$\hat{\mu}_e = \frac{1}{k}\sum_{r=1}^{k} |\hat{q}_r - v_r|$$

where k is the number of the elections, and $\hat{q}_r$ is defined by

$$\text{logit}(\hat{q}_r) = \text{logit}(\hat{v}_r) + \hat{\alpha}_r.$$

That is, to compute the election bias we set the poll-level bias component (γ_i) to zero.

To check that our Bayesian framework does not find substantial bias where none exists, we first fit the model to synthetic data generated via simple random sampling, preserving the empirically observed election outcomes, the number and date of polls in each election, and the size of each poll. On this synthetic dataset, we find both $\hat{\mu}_e$ and $\hat{\mu}_p$ to be at most 0.1 percentage points (i.e., less than one-tenth of one percentage point), nearly identical to the theoretically correct answer of zero.

Table 2 summarizes the results of fitting the Bayesian model on our primary polling dataset. (The full distribution of election-level estimates is provided in the Appendix.)

	Senatorial	Gubernatorial	Presidential
Average election-level absolute bias	2.0%	2.1%	1.0%
Average poll-level absolute bias	2.2%	2.3%	1.1%

Table 2: *Model-based estimates of election-level poll bias and variance, both of which are higher than would be expected from simple random sampling. If polls were simple random samples, average election-level standard error would be 1.9 percentage points, and bias would be zero.*

Consistent with our previous analysis, elections for all three offices exhibit substantial average absolute bias, approximately 2 percentage points for senatorial and gubernatorial elections and 1 percentage point for presidential elections. As expected, average absolute bias as estimated by the Bayesian model is somewhat smaller, and ostensibly more accurate, than what we obtained from the simple sample averages. However, we still find that poll bias is large; it is about as big as the theoretical sampling variation from SRS.

Why do polls exhibit nonnegligible election-level bias? We offer two possibilities. First, as discussed above, polls in a given election often have similar sampling frames. Telephone surveys, regardless of the organization that conducts them, will miss those who do not have a telephone. Relatedly, projections about who will vote—often based on standard likely voter screens—do not vary much from poll to poll, and as a consequence, election day surprises (e.g., an unexpectedly high number of minorities or young people turning out to vote) affect all polls similarly. Second, since polls often apply similar methods to correct for nonresponse, errors in these methods can again affect all polls in a systematic way. For example, it has recently been shown that supporters of the trailing candidate are less likely to respond to polls, even after adjusting for demographics [Gelman et al., 2016]. Since most polling organizations do not correct for such partisan selection effects, their polls are all likely to be systematically skewed.

Figure 4 shows how the average absolute election-level bias changes from one election cycle to the next. To estimate average absolute bias for each year, we average the estimated absolute election bias for all elections that year. While there is noticeable year-to-year

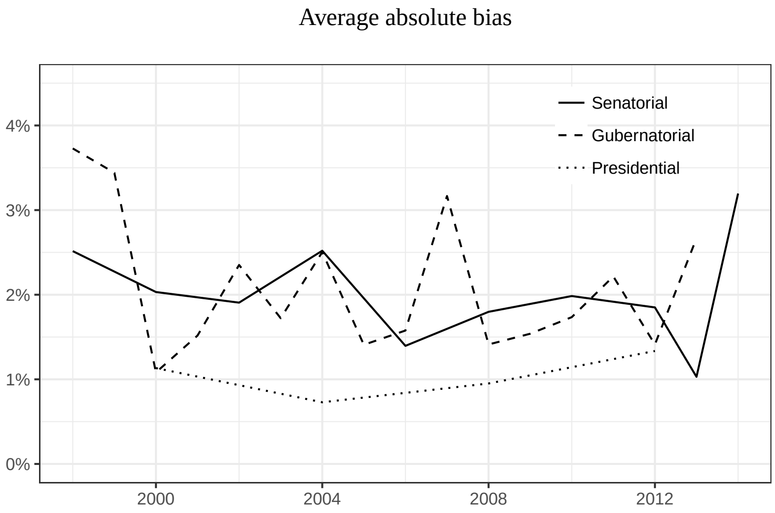

Figure 4: *Model-based estimates of average absolute bias show no consistent time trends across election cycles.*

variation, the magnitude is consistent over time, providing further evidence that the effects we observe are real and persistent. One might have expected to see a rise in poll bias over time given that survey response rates have plummeted—from an average of 36% in 1998 to 9% in 2012 [Pew Research Center, 2012]. One possibility is that pre- and post-survey adjustments to create demographically balanced samples mitigate the most serious issues associated with falling response rates, while doing little to correct for the much harder problem of uncertainty in turnout.

Finally, Figure 5 shows the relationship between election-level bias in elections for different offices within a state. Each point corresponds to a state, and the panels plot estimated bias for the two elections indicated on the axes. Overall, we find moderate correlation in bias for elections within the state: 0.41 for gubernatorial vs. senatorial, 0.49 for presidential vs. senatorial, and 0.37 for gubernatorial vs. presidential.[6] Such correlation again likely comes from a combination of frame and nonresponse errors. For example, since party-line voting is relatively common, an unusually high turnout of Democrats on election day could affect

[6] To calculate these numbers, we removed an extreme outlier that is not shown in Figure 3, which corresponds to polls conducted in Utah in 2004. There are only two polls in the dataset for each race in Utah in 2004.

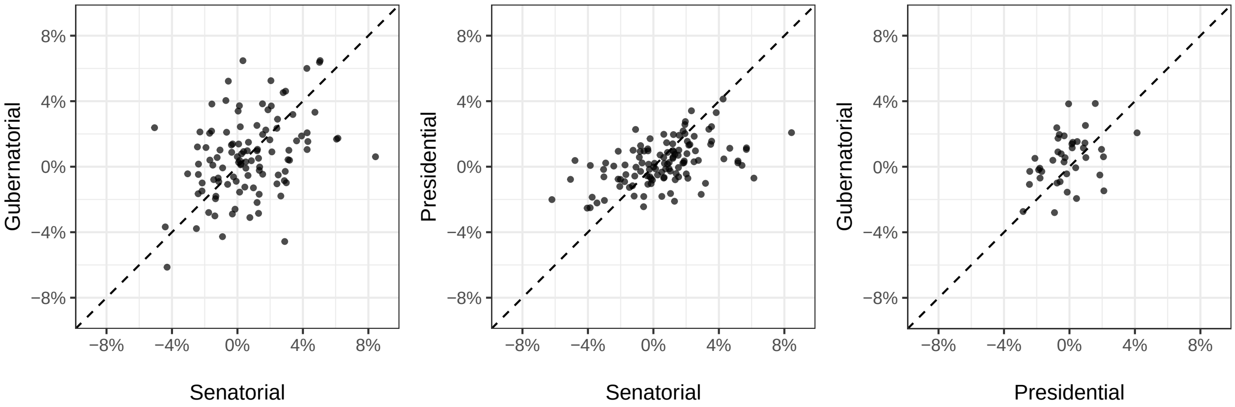

Figure 5: *Comparison of election-level polling bias in various pairs of state-level elections. Each point indicates the estimated bias in two different elections in the same state in the same year. The plots show modest correlations, suggesting a mix of frame and nonresponse errors.*

Priors		Senatorial	Gubernatorial	Presidential
$\mu_\alpha \sim N(0, 0.5)$,	election-level bias	2.2%	2.3%	1.1%
$\sigma_\alpha, \sigma_\gamma \sim N_+(0, 0.5)$	poll-level bias	2.0%	2.1%	1.0%
$\mu_\alpha \sim N(0, 0.05)$,	election-level bias	2.2%	2.3%	1.1%
$\sigma_\alpha, \sigma_\gamma \sim N_+(0, 0.05)$	poll-level bias	2.0%	2.1%	1.0%
$\mu_\alpha \sim N(0, 0.2)$,	election-level bias	2.2%	2.3%	1.1%
$\sigma_\alpha, \sigma_\gamma \sim$ inv-gamma$(3.6, 0.4)$	poll-level bias	2.0%	2.1%	1.0%

Table 3: *Posterior estimates for various choices of priors. Our results are nearly identical under various reasonable prior distributions, which makes sense given the large amount of data in our analysis.*

the accuracy of polling in multiple races. This correlated bias in turn leads to correlated errors, and illustrates the importance of treating polling results as correlated rather than independent samples of public sentiment.

We conclude our analysis by examining the robustness of our results to the choice of priors in the Bayesian model. In our primary analysis, the bias hyper-parameters μ_α has $N(0, \lambda)$ prior, and the variance hyper-parameters σ_α and σ_γ had $N_+(0, \lambda)$ priors, with $\lambda = 0.2$ in all cases. We consider three variations of this setup. First, we set $\lambda = 0.5$, corresponding to a prior that is effectively flat over the feasible parameter region. Second, we set $\lambda = 0.05$, corresponding to an informative prior that constrains the bias and excess variance to be relatively small. Finally, we replace the half-normal prior on the variance hyper-parameters with an inverse gamma distribution having parameters $\alpha = 3.6$ and $\beta = 0.4$; α and β were chosen so that the resulting distribution has mean and variance approximately equal to that of a $N_+(0, 0.2)$ distribution. Table 3 shows the results of this sensitivity analysis. Our posterior estimates are nearly identical in all cases, regardless of which priors are used.

5 Discussion

Researchers and practitioners have long known that traditional margins of error understate the uncertainty of election polls, but by how much has been hard to determine for two reasons. First, until recently it has been difficult to compile a large number of historical election polls to rigorously analyze. Second, estimating election-level bias and variance is a challenging statistical problem, since only a handful of polls are typically available in each race, and we must estimate quantities on the order of 1 percentage point. We address this second obstacle by developing a hierarchical Bayesian latent variable model that can accurately estimate the distribution of election-level bias and variance from small samples.

We estimate average absolute bias is 2.0 percentage points for senate races, 2.1 percentage points for gubernatorial races, and 1.0 percentage point for presidential races. At the very least, these findings suggest that care should be taken when using poll results to assess a candidate's reported lead in a competitive race. Moreover, in light of the correlated polling errors that we find, close poll results should give one pause not only for predicting the outcome of a single election, but also for predicting the collective outcome of related races. To mitigate the recognized uncertainty in any single poll, it has become increasingly common to turn to aggregated poll results, whose nominal variance is often temptingly small. While aggregating results is generally sensible, it is particularly important in this case to remember that shared election-level poll bias persists unchanged, even when averaging over a large number of surveys.

The 2016 U.S. presidential election offers a timely example of how correlated poll errors can lead to spurious predictions; see, for example, Gelman and Azari [2017]. Up through the final stretch of the campaign, nearly all pollsters declared Hillary Clinton the overwhelming favorite. The New York Times, for example, placed the probability of a Clinton win at 85% on the day before the election. Donald Trump ultimately lost the popular vote, but beat forecasts by about 2 percentage points. He ended up carrying nearly all the key swing states, including Florida, Iowa, Pennsylvania, Michigan, and Wisconsin, resulting in an electoral college win and the presidency. Because of shared poll bias—both for multiple polls forecasting the same state-level race, and also for polls in different states—even modest errors significantly impact a candidate's likelihood of winning. Such correlated errors might arise from a variety of sources, including frame errors due to incorrectly estimating the turnout population. For example, a higher-than-expected turnout among white men, or other Republican-leaning groups, may have skewed poll predictions across the nation.

Our analysis offers a starting point for polling organizations to quantify the uncertainty in predictions left unmeasured by traditional margins of errors. Instead of simply stating that these commonly reported metrics miss significant sources of error, which is the status quo, these organizations could—and, we feel, should—start quantifying and reporting the gap between theory and practice. Indeed, empirical election-level bias and variance could be directly incorporated into reported margins of error. Though it is hard to estimate these quantities for any particular election, historical averages could be used as proxies.

Large election-level bias does not afflict all estimated quantities equally. For example, it is common to track movements in sentiment over time, where the precise absolute level of support is not as important as the change in support. A stakeholder may primarily be interested in whether a candidate is on an up or downswing rather than his or her exact standing. In this case, the bias terms—if they are constant over time—cancel, and traditional methods may adequately capture poll error.

Given the considerable influence election polls have on campaign strategy, media narratives, and popular opinion, it is important to not only have accurate estimates of candidate support, but also accurate accounting of the error in those estimates. Looking forward, we hope our analysis and methodological approach provide a framework for understanding, incorporating, and reporting errors in election polls.

References

Stephen D. Ansolabehere and Thomas R. Belin. Poll faulting. *Chance*, 6(1):22–28, 1993.

Paul P. Biemer. Total survey error: Design, implementation, and evaluation. *Public Opinion Quarterly*, 74(5):817–848, 2010.

Gallup. Gallup world poll research design. http://media.gallup.com/WorldPoll/PDF/WPResearchDesign091007bleeds.pdf, 2007. Accessed: 2016-04-07.

Andrew Gelman. Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, 1(3):515–534, 2006.

Andrew Gelman and Julia Azari. 19 things we learned from the 2016 election. Technical report, Columbia University, Department of Political Science, 2017. URL http://www.stat.columbia.edu/~gelman/research/unpublished/what_learned_in_2016_4.pdf.

Andrew Gelman and Jennifer Hill. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press, 2007.

Andrew Gelman, Sharad Goel, Douglas Rivers, and David Rothschild. The mythical swing voter. *Quarterly Journal of Political Science*, 11:103–130, 2016.

Robert M. Groves and Lars Lyberg. Total survey error: Past, present, and future. *Public Opinion Quarterly*, 74(5):849–879, 2010.

Matthew D. Hoffman and Andrew Gelman. The no-U-turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, 15 (Apr):1593–1623, 2014.

Sharon Lohr. *Sampling: Design and Analysis, second edition.* Duxbury, 2009.

Sam G. McFarland. Effects of question order on survey responses. *Public Opinion Quarterly*, 45(2):208–215, 1981.

Pew Research Center. Assessing the representativeness of public opinion surveys. http://www.people-press.org/2012/05/15/assessing-the-representativeness-of-public-opinion-surveys, 2012. Accessed: 2016-04-07.

Pew Research Center. Our survey methodology in detail. http://www.people-press.org/methodology/our-survey-methodology-in-detail, 2016. Accessed: 2016-04-07.

Tom W. Smith. That which we call welfare by any other name would smell sweeter: An analysis of the impact of question wording on response patterns. *Public Opinion Quarterly*, 51(1):75–83, 1987.

Stan Development Team. Stan: A C++ library for probability and sampling, version 2.16.0, 2017. URL http://mc-stan.org/.

D. Stephen Voss, Andrew Gelman, and Gary King. Pre-election survey methodology: Details from nine polling organizations, 1988 and 1992. *Public Opinion Quarterly*, 59:98–132, 1995.

A Appendix

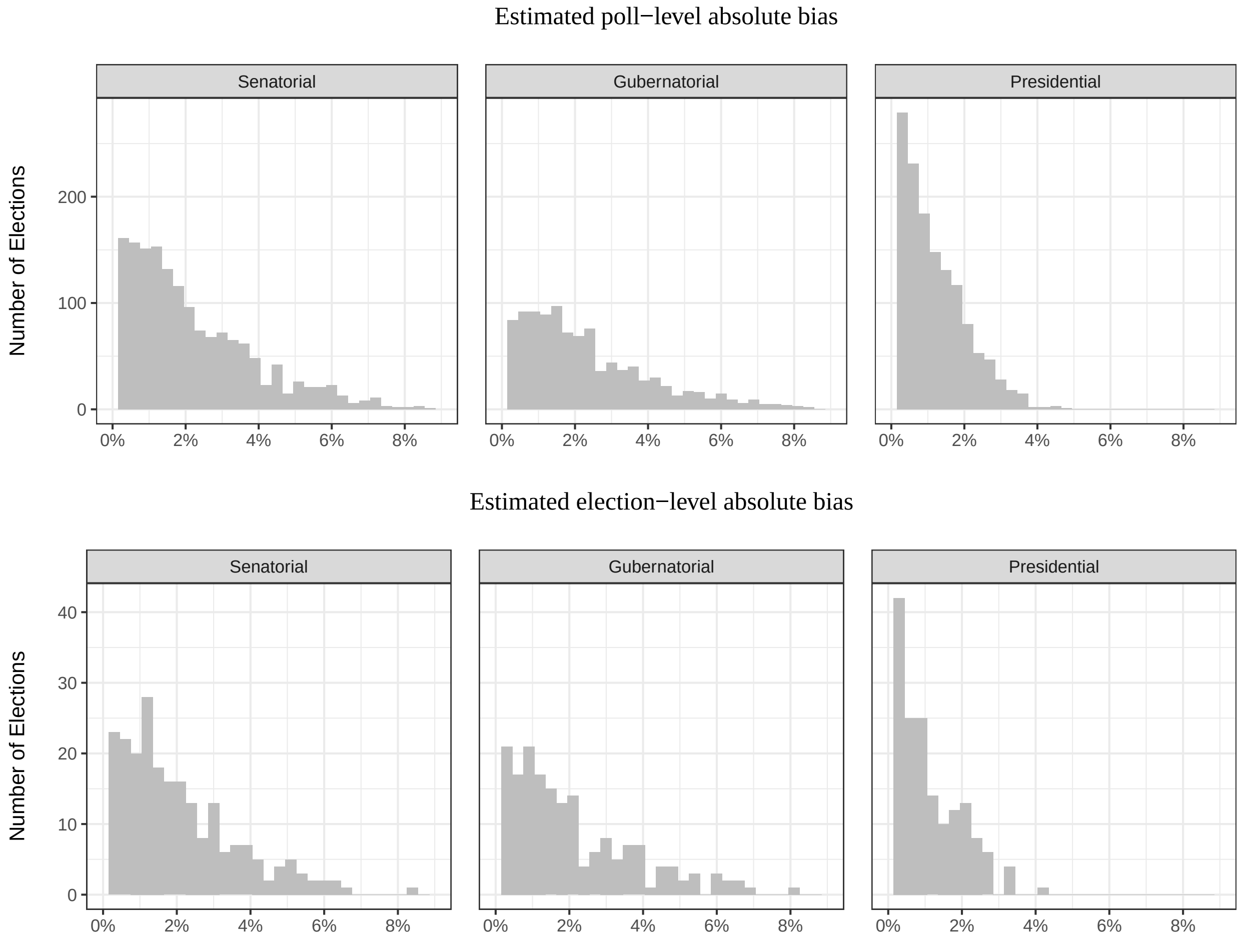

Figure 6: *Bayesian model estimates of election-level and poll-level absolute bias.*