Disentangling Bias and Variance in Election Polls

Houshmand Shirani-Mehr
Stanford University

David Rothschild
Microsoft Research

Sharad Goel
Stanford University

Andrew Gelman
Columbia University

February 3, 2018

Abstract

It is well known among researchers and practitioners that election polls suffer from a variety of sampling and non-sampling errors, often collectively referred to as *total survey error*. Reported margins of error typically only capture sampling variability, and in particular, generally ignore non-sampling 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 survey error into election-level bias and variance terms. We find that average absolute election-level bias is about 2 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 that they rely on similar screening rules when estimating who will vote. We also find that average election-level variance is higher than implied by simple random sampling, in part because polling organizations often use complex sampling designs and adjustment procedures. 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 the field: random sampling. As is recounted in so many textbooks, the huge but uncontrolled Literary Digest poll was trounced by Gallup's small, nimble random sample back in 1936. 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

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 both a point estimate of the election outcome, and also an estimate of the error in that prediction due to sample variance which accounts for the survey

[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].

weights [Lohr, 2009]. In practice, however, polling organizations often use the weights only in computing estimates, ignoring them when computing standard errors and instead reporting 95% margins of error based on the formula for simple random sampling (SRS)—for example ± 3.5 percentage points for an election survey with 800 people. Appropriate correction for the "design effect" corresponding to unequal weights would increase margins of error (see, for example, Mercer [2016]). The increase in margin of error depends on the poll, as some surveys have self-weighting designs (i.e., the sampling is constructed so that no weights are used in the analysis) while others weight on many factors. For some leading polls, standard errors should be increased by a factor of 30% to account for the weighting.[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

[2]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, a polling organization that explicitly notes that it inflates variance to adjust for the survey weights. Similarly, for a sampling of 36 state-level polls for the 2012 presidential election, only 9 reported higher-than-SRS margins of error. Complete survey weights are available for 21 ABC, CBS, and Gallup surveys conducted during the 2012 election and deposited into Roper Center's iPOLL. To account for the weights in these surveys, standard errors should on average be multiplied by 1.3 (with an interquartile range of 1.2 to 1.4 across the surveys), compared to the standard errors assuming simple random sampling.

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 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 carry out a Bayesian meta-analysis. We find that average absolute election-level bias is about 2 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—across the state.

2 Data

2.1 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 date during which it was in the field. We do not include House elections in our analysis since polling is only available for a small and non-representative subset of such races.

To construct this 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

[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, about half of our empirical estimate of RMSE. As discussed in Footnote 2, many polling organizations do not adjust for survey weights when computing uncertainty estimates, which in part explains this gap.

completed during the final three weeks of the campaigns. We augment these polls with the 67 corresponding ones for 2014 posted on Pollster.com, where for consistency with the FiveThirtyEight data, we consider only those completed in the last three weeks of the campaigns. In total, we end 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 primary dataset described above, we also consider 7,040 polls completed 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.

2.2 Data exploration

For each poll in our primary dataset (i.e., 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. We repeat this process separately for senatorial, gubernatorial, and presidential polls. For comparison, the dotted lines show the theoretical distribution of polling errors assuming simple random sampling (SRS). Specifically, for each senate poll i we first simulate an SRS 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

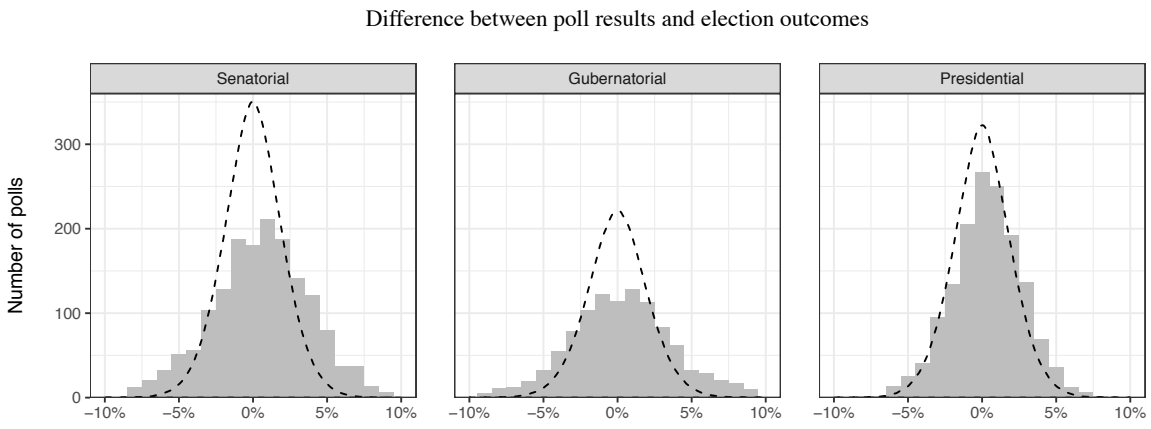

Figure 1: *The distribution of polling errors (Republican share of two-party support in the poll minus Republican share of the two-party vote in the election) for state-level presidential, senatorial, and gubernatorial election polls between 1998 and 2014. Positive values indicate the Republican candidate received more support in the poll than in the election. For comparison, the dashed lines show the theoretical distribution of polling errors assuming each poll is generated via simple random sampling.*

major-party candidates, and $v_{r[i]}$ is the final two-party vote share of the Republican candidate in the corresponding election $r[i]$. The dotted lines in the left-hand panel of Figure 1 show the distribution of errors across this set of synthetic senate polls. Theoretical SRS error distributions are generated analogously for 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 for any such aggregate bias. Second, the polls exhibit substantially larger errors than one would expect from SRS. 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

[4]Assuming N to be the number of polls, 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

senatorial and gubernatorial polls, in particular, have substantially larger RMSE (3.7% and 3.9%, respectively) than SRS (2.0% and 2.1%, respectively). In contrast, the RMSE for state-level presidential polls is 2.5%, not much larger than one would expect from SRS (2.0%). 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 is not immediately clear why presidential polls fare better, but one possibility is that turnout in such elections is easier to predict and so these polls suffer less from such error; in addition, presidential polls have higher visibility and so the organizations that conduct such surveys may invest more resources into their sampling and adjustment procedures.

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 non-sampling error, particularly frame and nonresponse error. As noted earlier, 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 possibilities. On one hand, election polls

in the corresponding election $r[i]$. Then RMSE is $\sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - v_{r[i]})^2}$.

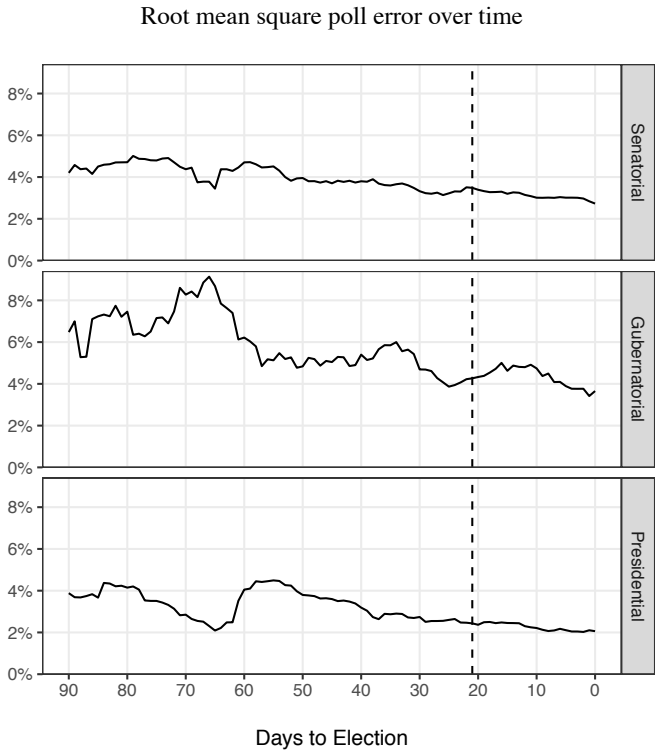

Figure 2: *Poll error, as measured by RMSE, over the course of elections. The RMSE on each day x indicates the average error for polls completed in a seven-day window centered at x. The dashed vertical line at the three-week mark shows that poll error is relatively stable during the final stretches of 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.*

may typically be unbiased but have large variance; on the other hand, polls in an election may generally have non-zero bias, but in aggregate these biases cancel to yield the depicted distribution. Our goal is to quantify the structure of polling errors. But before formally addressing this task we carry out the following simple analysis to build intuition. For each election r, we first compute the average poll estimate,

$$\bar{y}_r = \frac{1}{|S_r|} \sum_{i \in S_r} y_i,$$

where S_r is the set of polls in that election, and y_i is the two-party support for the Republican candidate in the i-th poll. Figure 3 (left) shows the difference between $\bar{y}_r$ and the

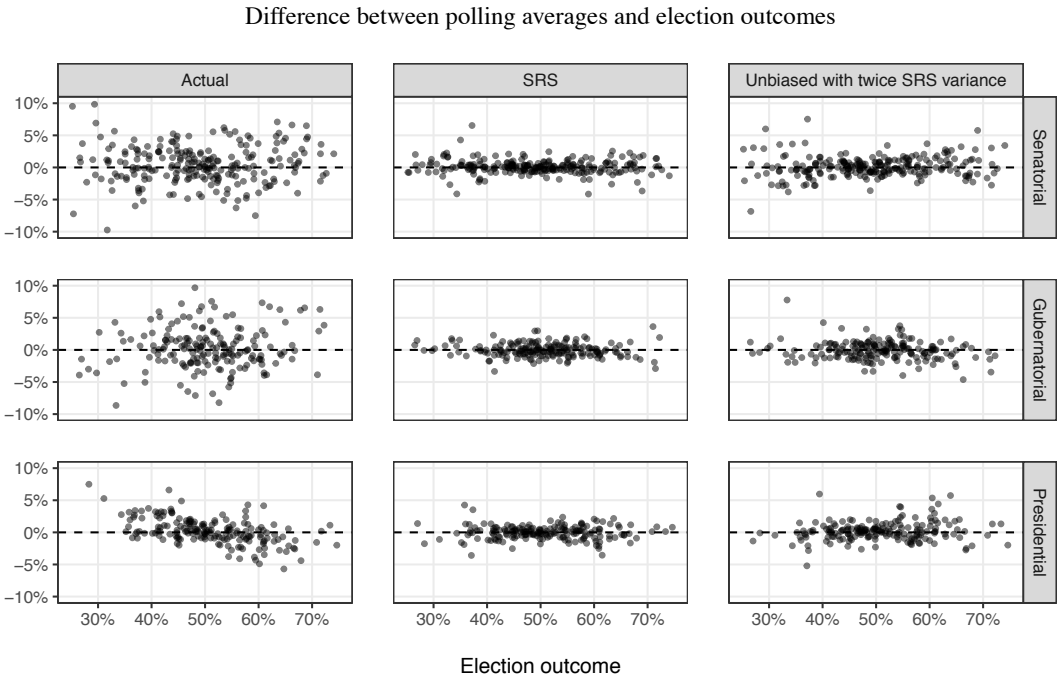

Figure 3: *Difference between polling averages and election outcomes (i.e., Republican share of the two-party vote), where each point is an election. The left panel shows results for the real polling data; the middle panel shows results for a synthetic dataset of SRS polls; and the right panel shows results for a synthetic dataset of polls that are unbiased but that have twice the variance of SRS.*

election outcome (i.e., the difference between the two-party poll average and the two-party Republican vote share), where each point in the plot is an election. For comparison, Figure 3 (middle) shows the same quantities for synthetic SRS polls, generated as above. It is visually apparent that the empirical poll averages are significantly more dispersed than expected under SRS. Whereas Figure 1 indicates that individual polls are over-dispersed, Figure 3 shows that poll averages also exhibit considerable over-dispersion. Finally, Figure 3 (right) plots results for synthetic polls that are unbiased but that have twice the variance as SRS. Specifically, we simulate a polling result by drawing a sample from a binomial distribution with parameters v_r (the election outcome) and $n_i/2$ (half the number of respondents in the real poll), since halving the size of the poll doubles the variance. Doubling poll variance increases the dispersion of poll averages, but it is again visually apparent that the empirical

poll averages are substantially more variable, particularly for senatorial and gubernatorial elections. Figure 3 shows that even a substantial amount of excess variance in polls cannot fully explain our empirical observations, and thus points to the importance of accounting for election-level bias.

3 A model for election polls

We now present and fit a statistical model to shed light on the structure of polling results. The bias term in our model captures systematic errors shared by all polls in an election (e.g., due to shared frame errors). The variance term captures residual dispersion, from traditional sampling variation as well as variation due to differing survey methodologies across polls and polling organizations. Our approach can be thought of as a Bayesian meta-analysis of survey results.

For each poll i in election $r[i]$ conducted at time t_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-party candidates. Let $v_{r[i]}$ denote the final two-party vote share for the Republican candidate. Then we model the poll outcome y_i as a random draw from a normal distribution parameterized as follows:

$$\begin{aligned} y_i &\sim \mathrm{N}(p_i\,,\,\sigma_i^2) \\ \mathrm{logit}(p_i) &= \mathrm{logit}(v_{r[i]}) + \alpha_{r[i]} + \beta_{r[i]} t_i \\ \sigma_i^2 &= \frac{p_i(1-p_i)}{n_i} + \tau_{r[i]}^2. \end{aligned} \tag{1}$$

Here, $\alpha_{r[i]} + \beta_{r[i]} t_i$ is the bias of the i-th poll (positive values indicate the poll is likely to overestimate support for the Republican candidate), where we allow the bias to change linearly over time.[5] The possibility of election-specific excess variance (relative to SRS) in

[5]To clarify our notation, we note that for each poll i, $r[i]$ denotes the election for which the poll was conducted, and $\alpha_{r[i]}$, $\beta_{r[i]}$, and $\tau_{r[i]}$ denote the corresponding coefficients for that election. Thus, for each

poll results is captured by the $\tau^2_{r[i]}$ term. Such excess variance may, for example, result from complex sampling designs and adjustment procedures. Estimating excess variance is statistically and computationally tricky, and there are many possible ways to model it. For simplicity, we use an additive term, and note that our final results are robust to natural alternatives; for example, we obtain qualitatively similar results if we assume a multiplicative relationship.

When modeling poll results in this way, one must decide which factors to include as affecting the mean p_i rather than the variance σ_i^2. For example, in our current formulation, systematic differences between polling firms [Silver, 2017] are not modeled as part of p_i, and so these "house effects" implicitly enter in the σ_i^2 term. There is thus no perfect separation between bias and variance, as explicitly accounting for more sources of variation when modeling the mean increases estimates of bias while simultaneously decreasing estimates of variance. Nevertheless, as our objective is to understand the election-level structure of polls, our decomposition above seems natural and useful.

To partially pool information across elections, we place a hierarchical structure on the parameters [Gelman and Hill, 2007]. We specifically set,

$$\alpha_j \sim \mathrm{N}(\mu_\alpha\,,\,\sigma_\alpha^2)$$
$$\beta_j \sim \mathrm{N}(\mu_\beta\,,\,\sigma_\beta^2)$$
$$\tau_j^2 \sim \mathrm{N}_+(0\,,\,\sigma_\tau^2)$$

where N_+ denotes the half-normal distribution. Finally, weakly informative priors are assigned to the hyper-paramaters μ_α, σ_α, μ_β, σ_β and σ_τ. Namely, $\mu_\alpha \sim \mathrm{N}(0, 0.2^2)$, $\sigma_\alpha \sim \mathrm{N}_+(0, 0.2^2)$, $\mu_\beta \sim \mathrm{N}(0, 0.2^2)$, $\sigma_\beta \sim \mathrm{N}_+(0, 0.2^2)$, and $\sigma_\tau \sim \mathrm{N}_+(0, 0.05^2)$. Our priors are weakly informative in that they allow for a large, but not extreme, range of parameter values. In particular, though a 5 percentage point (which is roughly equivalent to 0.2 on the logit scale)

election j, there is one $(\alpha_j, \beta_j, \tau_j)$ triple. Our model allows for a linear time trend (β_j) but we note that our empirical results are qualitatively similar even without this term.

poll bias or excess dispersion would be substantial, it is of approximately the right magnitude. We note that while an inverse gamma distribution is a traditional choice of prior for variance parameters, it rules out values near zero [Gelman et al., 2006]; our use of half-normal distributions is thus more consistent with our decision to select weakly informative priors. In Section 4.3, we experiment with alternative prior structures and show that our results are robust to the exact specification.

4 Results

4.1 Preliminaries

We fit the above model separately for senatorial, presidential and gubernatorial elections. 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.

The fitted model lets us estimate three key quantities. First, we define average election-level absolute bias μ_b by:

$$\mu_b = \frac{1}{k} \sum_{r=1}^{k} |b_r|$$

where k is the total number of elections in consideration (across all years and states), and b_r is the bias for election r. Specifically, b_r is defined by

$$b_r = \frac{1}{|S_r|} \sum_{i \in S_r} (p_i - v_r)$$

where S_r is the set of polls in election r, and p_i is computed via Eq. (1). That is, to compute b_r we average the bias for each poll in the election. Second, we define the average absolute bias on election day μ_{b_0} by:

$$\mu_{b_0} = \frac{1}{k} \sum_{r=1}^{k} |q_r - v_r|,$$

	Senatorial	Gubernatorial	Presidential
Average election-level absolute bias ($\hat{\mu}_b$)	2.1% (0.10%)	2.3% (0.10%)	1.2% (0.07%)
Average election-level absolute bias on election day ($\hat{\mu}_{b_0}$)	2.0% (0.13%)	2.2% (0.12%)	1.2% (0.08%)
Average election-level standard deviation ($\hat{\mu}_\sigma$)	2.8% (0.07%)	2.7% (0.07%)	2.2% (0.04%)

Table 1: *Mean posterior estimates of election-level poll bias and standard deviation, with the standard deviation of the posterior distribution given in parentheses. Bias and standard deviation are higher than would be expected from SRS. Under SRS, the average election-level standard deviation would be 2.0 percentage points for senatorial and presidential polls, and 2.1 percentage points for gubernatorial polls; the bias would be zero.*

where q_r is defined by

$$\text{logit}(q_r) = \text{logit}(v_r) + \alpha_r.$$

That is, we define μ_{b_0} by setting t_i to zero in Eq. (1). Finally, we define average election-level standard deviation μ_σ by:

$$\mu_\sigma = \frac{1}{k}\sum_{r=1}^{k} \sigma_r$$

where

$$\sigma_r = \frac{1}{|S_r|}\sum_{i \in S_r} \sigma_i.$$

To check that our modeling framework produces accurate estimates, we first fit it on synthetic data generated via SRS, 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 the mean posterior estimates $\hat{\mu}_b$ and $\hat{\mu}_{b_0}$ are approximately 0.2 percentage points (i.e., approximately two-tenths of one percentage point), nearly identical to the theoretically correct answer of zero. We further find that the posterior mean $\hat{\mu}_\sigma$ is approximately 2.1 percentage points, closely aligned with the theoretically correct answer of 2.0.

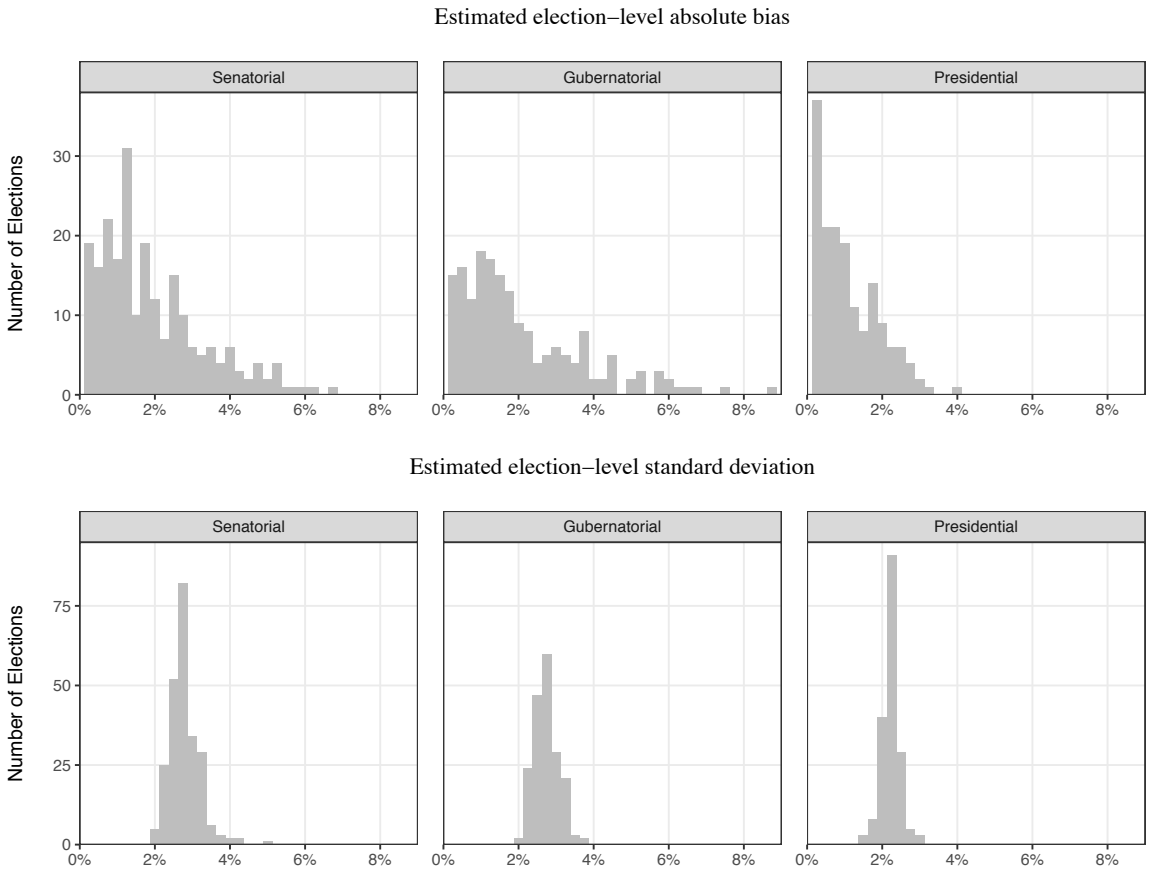

Figure 4: *Model estimates of election-level absolute bias (top plots) and election-level standard deviation (bottom plots).*

4.2 Empirical results

Table 1 summarizes the results of fitting the model on our primary polling dataset. The results show elections for all three offices exhibit substantial average election-level absolute bias, approximately 2 percentage points for senatorial and gubernatorial elections and 1 percentage point for presidential elections. The poll bias is about as big as the theoretical sampling variation from SRS. The full distribution of election-level mean posterior estimates is shown in Figure 4. The top panel in the plot shows the distribution of $|\hat{b}_r|$, and the bottom panel shows $\hat{\sigma}_r$.

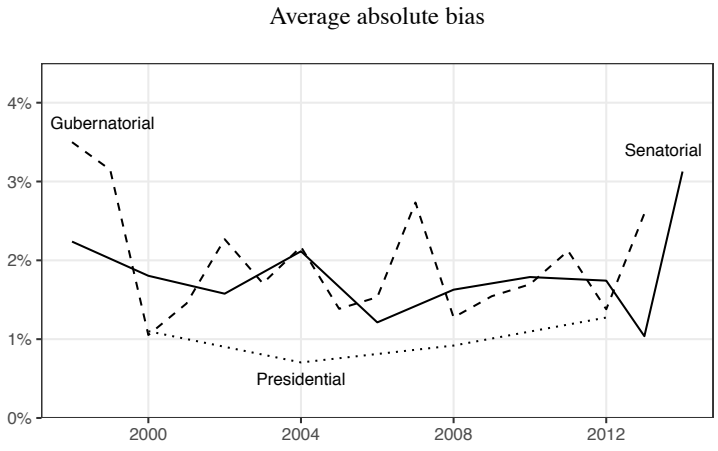

Figure 5: *Model-based estimates of average absolute bias show no consistent time trends across election cycles.*

Why do polls exhibit non-negligible election-level bias? We offer three possibilities. First, as discussed above, polls in a given election often have similar sampling frames. As an extreme example, telephone surveys, regardless of the organization that conducts them, will miss those who do not have a telephone. More generally, polling organizations are likely to undercount similar, hard-to-reach groups of people (though post-sampling adjustments can in part correct for this). 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. Finally, respondents might misreport their vote intentions, perhaps because of social desirability bias (if they support a polarizing candidate) or acquiescence bias (if they believe the poll to be leaning against their preferred candidate).

Figure 5 shows how the average absolute election-level bias changes from one election

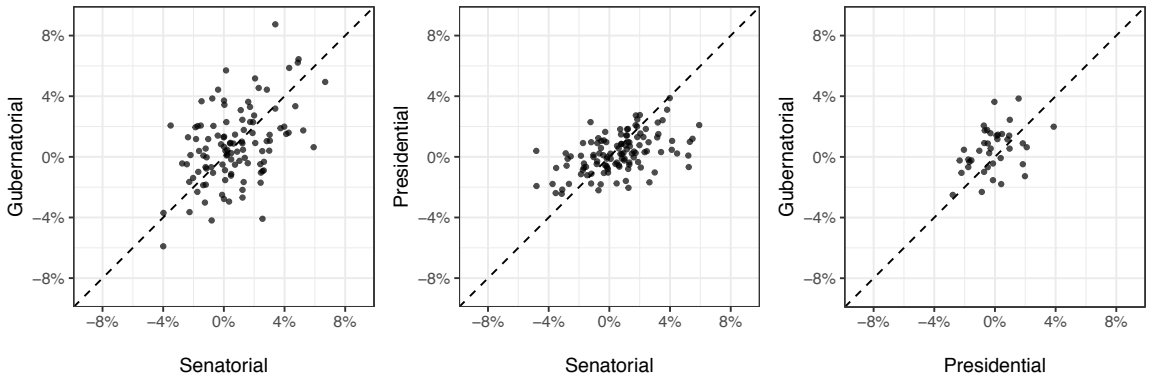

Figure 6: *Comparison of election-level polling bias in various pairs of state-level elections. Each point indicates the estimated bias in two different elections in the same state in the same year. The plots show modest correlations, suggesting a mix of frame and nonresponse errors.*

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 variation, the magnitude is consistent over time, providing further evidence that the effects we observe are real and persistent. We note that 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 6 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: 0.45 for gubernatorial vs. senatorial, 0.50 for presidential vs. senatorial, and 0.39 for gubernatorial vs. presidential.[6] Such correlation again likely comes from a combination of frame and nonresponse errors. For example, since party-line voting

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

is relatively common, an unusually high turnout of Democrats on election day could affect 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.

4.3 Sensitivity analysis

We conclude our analysis by examining the robustness of our results to the choice of priors in the model. In our primary analysis, we consider a 5 percentage point (equivalent to 0.2 on the logit scale) standard deviation for the bias and variance hyper-parameters. In this section, we consider three alternative choices. First, we change the standard deviation defined for all hyper-parameters to approximately 25 percentage points, corresponding to a prior that is effectively flat over the feasible parameter region. Second, we change the standard deviation to approximately one percentage point, 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; α and β were chosen so that the resulting distribution has mean and variance approximately equal to that of the half-normal distribution in the original setting. Table 2 shows the results of this sensitivity analysis. Our posterior estimates are qualitatively similar in all cases, regardless of which priors are used. While the posterior estimates for absolute bias are nearly identical, inverse gamma priors on the variance hyper-parameters result in somewhat higher estimated election-level standard deviation.

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, in part because of a lack of data. By compiling and analyzing a large collection of historical election

Priors	Measure	Sen.	Gov.	Pres.
$\mu_\alpha, \mu_\beta \sim \mathrm{N}(0, 1^2)$	absolute bias	2.1%	2.3%	1.2%
$\sigma_\alpha, \sigma_\beta \sim \mathrm{N}_+(0, 1^2)$	election day absolute bias	2.0%	2.2%	1.2%
$\sigma_\tau \sim \mathrm{N}_+(0, 0.2^2)$	standard deviation	2.8%	2.7%	2.2%
$\mu_\alpha, \mu_\beta \sim \mathrm{N}(0, 0.04^2)$	absolute bias	2.0%	2.3%	1.2%
$\sigma_\alpha, \sigma_\beta \sim \mathrm{N}_+(0, 0.04^2)$	election day absolute bias	2.0%	2.2%	1.2%
$\sigma_\tau \sim \mathrm{N}_+(0, 0.01^2)$	standard deviation	2.8%	2.7%	2.2%
$\mu_\alpha, \mu_\beta \sim \mathrm{N}(0, 0.2^2)$	absolute bias	1.9%	2.1%	1.1%
$\sigma_\alpha, \sigma_\beta \sim \mathrm{Gamma}^{-1}(3.6, 0.4)$	election day absolute bias	1.8%	2.0%	1.0%
$\sigma_\tau \sim \mathrm{Gamma}^{-1}(3.6, 0.1)$	standard deviation	3.3%	3.4%	2.9%

Table 2: *Mean posterior estimates for various choices of priors. Our results are qualitatively similar regardless of the priors selected.*

polls, we find substantial election-level bias and excess variance. We estimate average absolute bias is 2.1 percentage points for senate races, 2.3 percentage points for gubernatorial races, and 1.2 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. Up through the final stretch of the campaign, nearly all pollsters declared Hillary Clinton the overwhelming favorite to win the election. 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 win estimates. 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 out.

Given the considerable influence election polls have on campaign strategy, media narratives, and popular opinion, it is important to have both accurate estimates of candidate support and also accurate accounting of the uncertainty in those estimates. Looking forward, we hope our analysis and methodological approach provide a framework for understanding, incorporating, and reporting errors in election polls.

Acknowledgement

The survey weights discussed in Footnote 2 are based on polls obtained from the iPOLL Databank provided by the Roper Center for Public Opinion Research at Cornell University. The data and code to replicate our results are available online at `https://github.com/5harad/polling-errors`.

References

Stephen Ansolabehere and Thomas R. Belin. Poll faulting. *Chance*, 6, 1993.

Paul P. Biemer. Total survey error: Design, implementation, and evaluation. *Public Opinion Quarterly*, 74(5):817–848, 2010. ISSN 0033-362X.

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

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*, 2016.

Andrew Gelman et al. Prior distributions for variance parameters in hierarchical models (comment on article by browne and draper). *Bayesian analysis*, 1(3):515–534, 2006.

Robert M. Groves and Lars Lyberg. Total survey error: Past, present, and future. *Public Opinion Quarterly*, 74(5):849–879, 2010. ISSN 0033-362X.

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*. Nelson Education, 2009.

Hao Lu and Andrew Gelman. A method for estimating design-based sampling variances for surveys with weighting, poststratification, and raking. *Journal of Official Statistics*, 19 (2):133, 2003.

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

Andrew Mercer. 5 key things to know about the margin of error in election polls. http://www.pewresearch.org/fact-tank/2016/09/08/understanding-the-margin-of-error-in-election-polls/, 2016. Pew Research Center.

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.

Nate Silver. Fivethirtyeight's pollster ratings, 2017. URL https://projects.fivethirtyeight.com/pollster-ratings/.

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.

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.