Disentangling *Disentangling Bias and Variance in Election Polls*: A Reproduction and Critical Evaluation

Applied Linear Regression

Wayne T. Lee

18 December 2022

Andersen Gu

Introduction:

This report attempts to reproduce and evaluate the Bayesian meta-analysis from *Disentangling Bias and Variance in Election Polls*, an article focusing on the decomposition of total survey error among state-level senatorial, gubernatorial, and presidential election polls. While the original study provides a variety of different metrics to gauge and differentiate total survey error into bias and sampling variance, this report reproduces the original paper's estimate of average election-level absolute bias, μ_b. Average election-level absolute bias was chosen given its qualitative significance as the main component of systematic election-level error among polling firms. Subsequently, evaluations and sensitivity analyses were conducted both on the dataset used and regarding the model itself. First, Chow tests were conducted to determine whether the assumed date by which the time-dependent component of bias can be assumed to be negligible was statistically substantiated. This report then compares the 95% confidence intervals of the posterior estimates of the time-dependent bias term between the original three week time window model and a model inclusive of poll results 100 days prior to election date. Second, assumptions regarding the additive nature of excess variance were tested by comparing the average election-level absolute bias and election-level bias estimates of the original model with those of an evaluative model with multiplicative excess variance. The posterior distributions of the aforementioned metrics were then bootstrapped to determine whether or not they were produced significantly different results. Ultimately, such sensitivity analyses conclude that, while certain assumptions made by the authors are arbitrary, the presented findings and models are generally robust to alterations both in the dataset and the structure of the model.

Background (Summary of the Original Paper):

The original paper attempts to decompose total survey error into distinct bias and sampling variance components through a Bayesian meta-analysis of existing polling data. More specifically, the paper attempts to identify whether there exists systematic election-level bias among polls distinct from inherent sampling variation. Gelman points out that most survey organizations underreport polling margins of errors by relying on false simple random sample ("SRS") assumptions that are not reflective of actual sampling practices among most polling firms. In practice, assumptions regarding SRS are false due to 1) shared bias among all election-level polls born from fundamental systematic errors such as frame errors in which polls

fail to derive a nationally representative sample (i.e. telephone poll only measures the opinion of those with telephones), and 2) non-uniform sample methodologies among different polling firms. The authors were able to obtain posterior point estimates of key metrics such as average election-level absolute bias, among others, via Hamiltonian Monte Carlo as implemented in Stan, a statistical language focusing on Bayesian inference. The presented model was inputted into Stan and fitted onto an original dataset compiled from various polling websites and databases (specified later in this report). Estimates for average election-level absolute bias were obtained with the senatorial average election-level absolute bias being 2.1%, gubernatorial average election-level absolute bias being 2.3%, and presidential election-level bias being 1.2%.

All statistical analyses in the original paper were based on a novel dataset consisting of 4221 polls for 608 state-level gubernatorial, senatorial, and presidential elections between 1998 to 2014. Poll results for each poll were recorded as the total Republican Party vote share among a two-party vote with bounds of 0 and 1 with 0 representing 0% of voters answering in favor of the Republican Party and 1 representing 100% Republican vote share. RMSE calculations for the initial data exploration was calculated by subtracting poll Republican vote share with the actual election Republican party vote share. Individual election poll data was collected and available by survey organizations such as FiveThirtyEight, Pollster.com, and RealClearPolitics.com. An additional auxiliary dataset consisting of 7040 polls from the last 100 days of 314 state-level senatorial, gubernatorial, and presidential elections was compiled from Pollster.com and RealClearPolitics.com. The auxiliary dataset was used in the paper to substantiate the assumption that the root mean square error ("RMSE") of all election polls under analysis vary negligibly within 21 days of election day.

While the presented model was comprehensive in its decomposition of total survey error, several assumptions remain open to evaluation and sensitivity analyses. First, the authors assume that the RMSEs of all election polls remain mostly stable, and thus time-independent, within three weeks of the election. While such an assumption allows the original authors to stray away from further complex time-dependent models, it was not fully substantiated through statistical analysis. Furthermore, when decomposing total survey error into bias and sampling variance, the model assumes an additive excess variance term, which represents variance born from complex and non-uniform sampling methodologies among different polling firms. The authors stated that the choice of an additive error term was chosen due to ease of computation and calculation in

addition to being qualitatively similar to a multiplicative error, but such an assumption is unsubstantiated in the statistical analysis within the paper itself. As such, this report will attempt to test both the time-independence and additive error assumptions in the sections below.

Presented Model:

The specific hierarchical model employed was parametrized as follows:

$$y_i \sim N(p_i, \sigma_i^2)$$

$$logit(p_i) = logit(\nu_{r[i]}) + \alpha_{r[i]} + \beta_{r[i]} t_i$$

$$\sigma_i^2 = \frac{p_i(1-p_i)}{n_i} + \tau_{r[i]}^2$$

For poll *i* in election *r[i]* in time t_i, y_i represents the poll's predicted total vote share for the Republican candidate, $\nu_{r[i]}$ represents the actual total vote share for the Republican candidate, and n_i represents the number of respondents. The $\alpha_{r[i]} + \beta_{r[i]} t_i$ term denotes systematic bias with the first element, $\alpha_{r[i]}$, corresponding to the time-independent component of bias and the second element, $\beta_{r[i]} t_i$, corresponding to the time-dependent component of bias. $\tau_{r[i]}^2$ represents election-specific excess variance due to complex sampling design that is not attributable to any systematic, election-level bias that may affect all polls predicting a specific election. The parameters were assigned hyperparameter which are listed as follows:

$$\alpha_j \sim N(\mu_\alpha, \sigma_\alpha^2)$$

$$\beta_j \sim N(\mu_\beta, \sigma_\beta^2)$$

$$\tau_j^2 \sim N_+(0, \sigma_\tau^2)$$

μ_α, μ_β, σ_α, σ_β, and σ_τ are all hyperparameters which themselves were assigned weakly informative hyperpriors of $\mu_\alpha \sim N(0, 0.2^2)$, $\sigma_\alpha \sim N_+(0, 0.2^2)$, $\mu_\beta \sim N(0, 0.2^2)$, $\sigma_\beta \sim N_+(0, 0.2^2)$, and $\sigma_\tau \sim N_+(0, 0.05^2)$. Of principal importance to this report, draws from the posterior distribution can be used to generate point estimates of average election-level absolute bias, μ_b, through the following equation:

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

With k corresponding to the total number of elections in consideration, and b_r representing the bias for election r. b_r is further modeled by the following equation with S_r representing the total number of polls for election r, p_i derived from the aforementioned logit model, and ν_r being derived from actual vote share for election *r*:

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

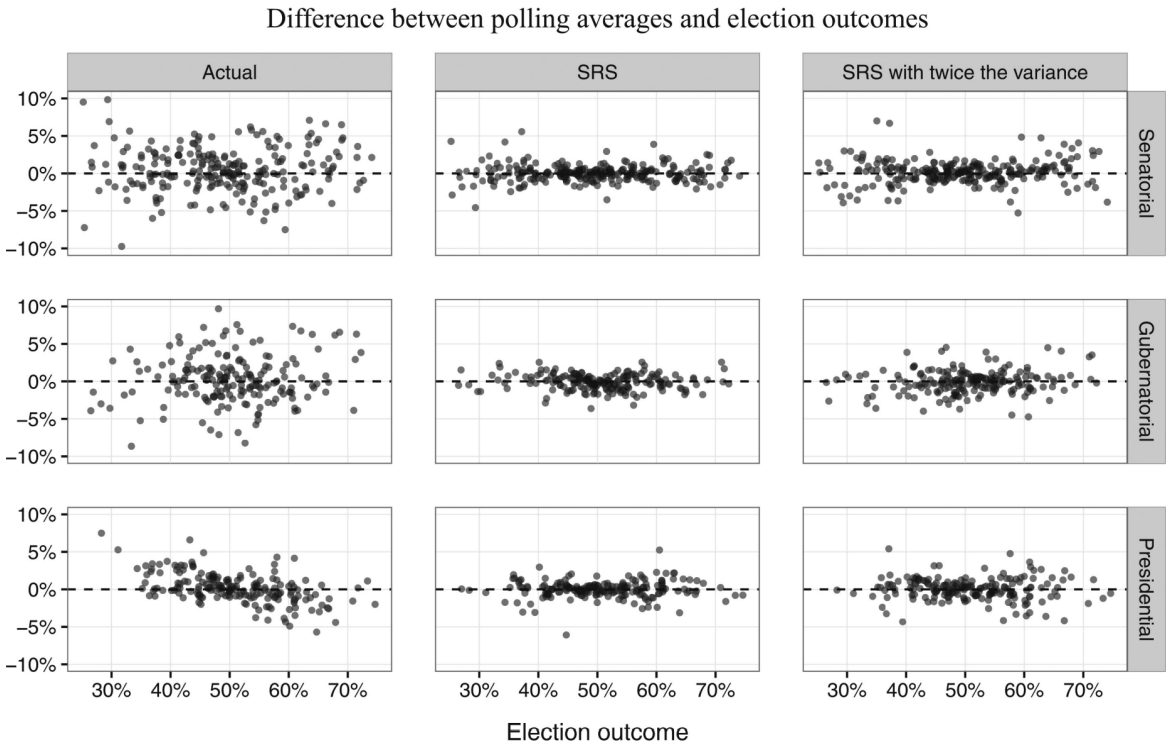
Such a model addresses the research question by providing a model to be fitted onto the primary dataset to derive point estimates of the average election-level absolute bias through simulated draws on the posterior distributions within the aforementioned hierarchical model. The hierarchical structure of the model also helps answer the research question by providing a basis for comparison among the thousands of polls under study. More specifically, by including a set of hyperparameters, the model implicitly asserts that the results of every poll can be characterized as following the same set of distributions. Furthermore, by pooling the information of all the polls under analysis, estimates of the bias on one poll can borrow strength from the others.

Description of the Data:

The dataset used in this report mirror those that were used in the original report. Namely, the primary dataset of 4221 polls will be used to reproduce the average election-level bias estimate generated by the authors with no alterations to the dataset and the model. Subsequently, after reproducing the original average election-level bias term, the auxiliary pollset is used to generate posterior estimates of the time-dependent bias term, β, between the original model inclusive of polls only within three weeks of election day and a model inclusive of polls within 100 days of election day. The auxiliary dataset was used given that the primary dataset was altered to exclude polls from outside the three week window from an original dataset compiled by the authors that was not provided in the replication materials. As such, while the auxiliary dataset provides a sufficient number of polls to generate meaningful posterior estimates of

average election-level bias and estimate the time-dependent bias term, an ideal dataset would have combined both primary and auxiliary datasets inclusive of polls conducted at least 100 days prior to election day. While analysis of such a dataset may require more intensive modelling, such data would be critical in verifying the assumptions of time negligibility which may provide evidence against competing time-dependent models such as those based on late swing paradigms and conditional non-response error (i.e. supporters of a trailing party are less likely to respond to polls if they believe their party will lose, an effect amplified closer to election day). In addition, while the primary dataset compiled by the authors is comprehensive, an ideal dataset may draw polling results from more databases other than the ones listed. By drawing from a limited number of sources, an extra layer of potential sampling frame error is introduced as each database itself does not contain an complete list of all available polls and can be thought of as itself sampling from the total population of polls. As such, conclusions should implicitly recognize the added layer of uncertainty in determining whether such systematic bias is commonplace for all polls or only endemic to the specific polling sample constituted by the dataset.

The graph plots below illustrate the aforementioned discrepancy between the distribution of empirical poll RMSEs and the distribution of hypothetical SRS-generated poll RMSEs, and helps distinguish between election-level bias, μ_b, and inherent sampling variance, σ_i^2.



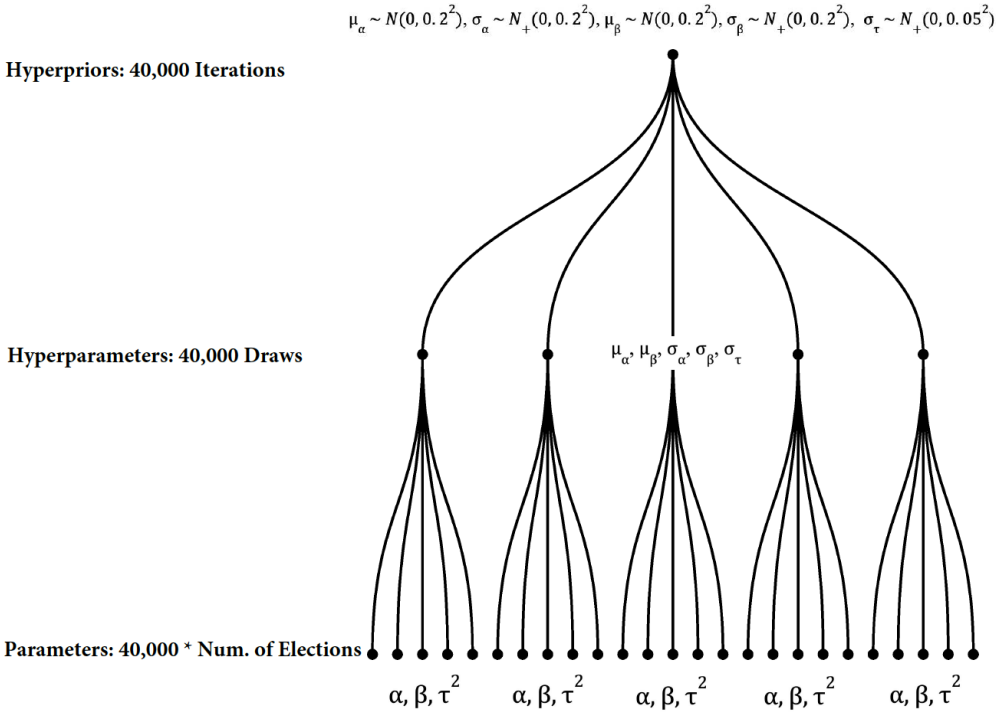
The graphs above plot the difference between election-level average polling prediction for given election *r[i]* and the actual two-party vote share. Each point represents a single election as opposed to any individual poll. Each hypothetical SRS polling result was generated based on a corresponding poll *i* by drawing a sample from a binomial distribution with parameters n_i, which represents the number of respondents in actual poll *i*, and $v_{r[i]}$, which represents the actual two-party vote share for election *r[i]* (multiple polls for the same election are compiled so SRS replications of such polls will have the same $v_{r[i]}$ but different n_i values). The average predicted vote share of the SRS simulated polls was then subtracted from the actual two-party vote share to generate each point on the SRS graphs. While it is often difficult to discern whether total survey error is due to purely bias or inherent sampling variance, the comparison of these three graphs helps illustrate the key variables of interest with the spread of the distribution of empirical average election-level bias being far greater than both the SRS distribution and the SRS with double variance distribution. This suggests that the empirical distribution could not have been generated purely by scaling the inherent sampling variance, σ_i^2, (which is represented by the third set of plots) and that such variation may be due to the other component of total survey error: election-level polling bias ($\alpha_{r[i]} + \beta_{r[i]} t_i$).

Reproduction of Results:

As stated before, the principal metric of focus is average election-level absolute bias represented as μ_b. Such a metric was chosen based on its importance in substantiating the authors' qualitative theories regarding systematic frame errors among polling firms and its interpretability in comparison to more abstract metrics such as average excess sampling variance. In preparing the data for statistical analysis, the primary dataset was first pre-processed with corrupted entries, polls conducted outside three weeks of election date, national-level polls, and House of Representatives polls removed. Polls outside three weeks of the election date were removed due to the assumption that changing voter attitudes may be more influential in determining poll predictions earlier in the campaign session which may obscure attempts to measure the effect of systematic frame errors. National-level polls were removed as national polls of senatorial and gubernatorial elections are generally non-representative of in-state voter attitudes. House of Representatives election polls were removed as polling data was available

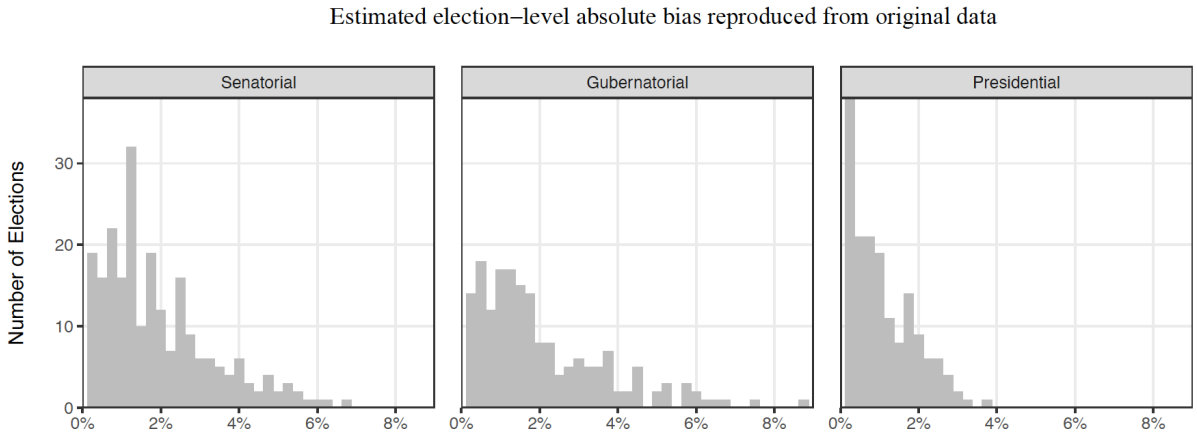
only to a small and non-representative subset of House polls. The cleaned data was then formatted and fitted via a pre-written Stan file that contained the outlined model above.

For each election-level, 40,000 iterations were run with the initial hyperprior distributions combined with the primary dataset to yield a subsequent posterior hyperprior distribution used as the prior for the next iteration. As such, each iteration entailed one draw from the resultant posterior distribution for a total of 40,000 sets of hyperparameter estimates (μ_α, μ_β, σ_α, σ_β, σ_τ). Each set of hyperparameter estimates then yielded 608 total sets of parameter estimates (α, β, τ^2) across all election-levels with each set of parameter estimates corresponds to each election *r[i]*. Each set of parameter estimates was then used to calculate p_i which was used to derive election-level absolute bias, b_r, which itself can be averaged to obtain average election-level absolute bias, μ_b. The tree diagram below illustrates the basic structure of the hierarchical model with each hyperparameter node representing one draw from the hyperprior distributions and each parameter node representing one set of parameters for one election:



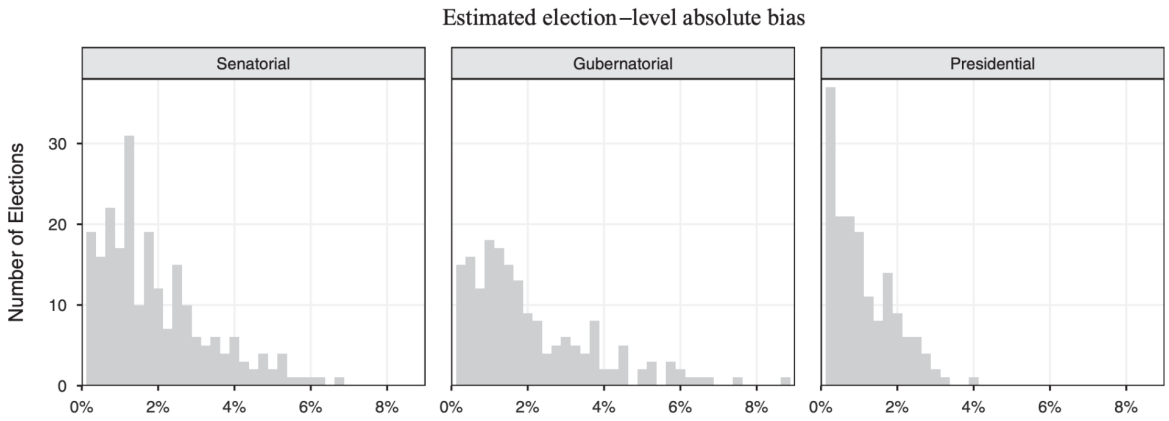
For example, given that there were 179 presidential polls in the primary dataset used, the total number of sets of parameter estimates was equal to 7,160,000 with each of the 40,000

iterations and draws producing 179 sets of parameter estimates. Point estimates were then generated by grouping the total 7,160,000 sets of parameters by their respective election and then taking the average. More specifically the parameter estimates were then inputted into the model presented above to calculate p_i which was then used to calculate each election's corresponding election-level bias, b_r. 7,160,000 b_r were produced which was subsequently divided by 40,000 (total number of iterations corresponding to each election). $|S_r|$ was implicitly assumed to be equivalent to the total number of polls for election *r[i]*. Such an operation left us with 179 point estimates of presidential election-level absolute bias, b_r, with the outlined process above repeated for senatorial and gubernatorial elections. The graph below shows the distribution of election-level absolute bias, b_r, for each race within the three election-levels in the primary dataset:



Averaging each distribution then yields the average election-level absolute bias which aligned almost exactly with the original paper's results. The calculated average election-level absolute bias for each election level is as follows: 2.1% for senatorial polls, 2.3% for gubernatorial polls, and 1.2% for presidential polls. In comparison, the average election-level absolute bias for each election-level was reported in the original article to be 2.1% for senatorial polls, 2.3% for gubernatorial polls, and 1.2% for presidential polls. this report's almost exact replication of the original paper's results is unsurprising Given the large number of polls used to

fit the model and the large number of iterations/draws. Small discrepancies exist between overall distribution of the election-level absolute biases with the the deviation of the posterior distribution for senatorial polling bias being 0.095 as opposed to the reported 0.1 in the original paper. The output from the original paper is posted below for reference:



Evaluations of the Model and Sensitivity Analyses:

Stress tests were conducted to test the fundamental assumptions of the original paper. In particular, this report conducted two sensitivity analyses testing 1) the assumption that the time-dependent component of bias, β, is negligible within three-weeks of election date, and 2) that the excess variance term is additive. The first sensitivity test regarding the time-dependent component of bias involves altering the dataset and interpreting the fitted model results to see if they align with the results of the original analysis. The second sensitivity test regarding excess variance involves altering the model and changing $\sigma_i^2 = \frac{p_i(1-p_i)}{n_i} + \tau_{r[i]}^2$ to $\sigma_i^2 = (\frac{p_i(1-p_i)}{n_i}) * \tau_{r[i]}^2$.

Time-Dependent Component of Bias:

With regard to the time-dependent component of bias, the original study had validated their assumption by using the auxiliary dataset to graph changes in average RMSE based on days from election date, and concluded that the RMSE of polls conducted closer to election day generally stabilize. However, no statistical analyses were conducted to substantiate their assumption with the three-week window mark set arbitrarily. To test whether the arbitrary

boundary is meaningful, this report modeled the arbitrary bound as the breakpoint in a piecewise function and examined whether the lines before and after the breakpoint are statistically significantly different. More specifically, Chow tests were conducted to test whether there existed a statistically significant difference between the true coefficients of two linear regressions. Qualitatively, the null hypothesis suggests that there exists no statistically significant difference between the two linear regression coefficients, which implies that the assumed three-week window does not mark a significant plateau in the rate of change of RMSE. As such, time-dependent bias may not be negligible if the null hypothesis were not rejected. Total survey error, represented as RMSE, was regressed on time in both sets of linear regressions, producing the basic linear models with X_1 representing days within the three-week window and X_2 representing days outside of the window:

$$RMSE_{before} = \widehat{time}_{before} * X_1$$

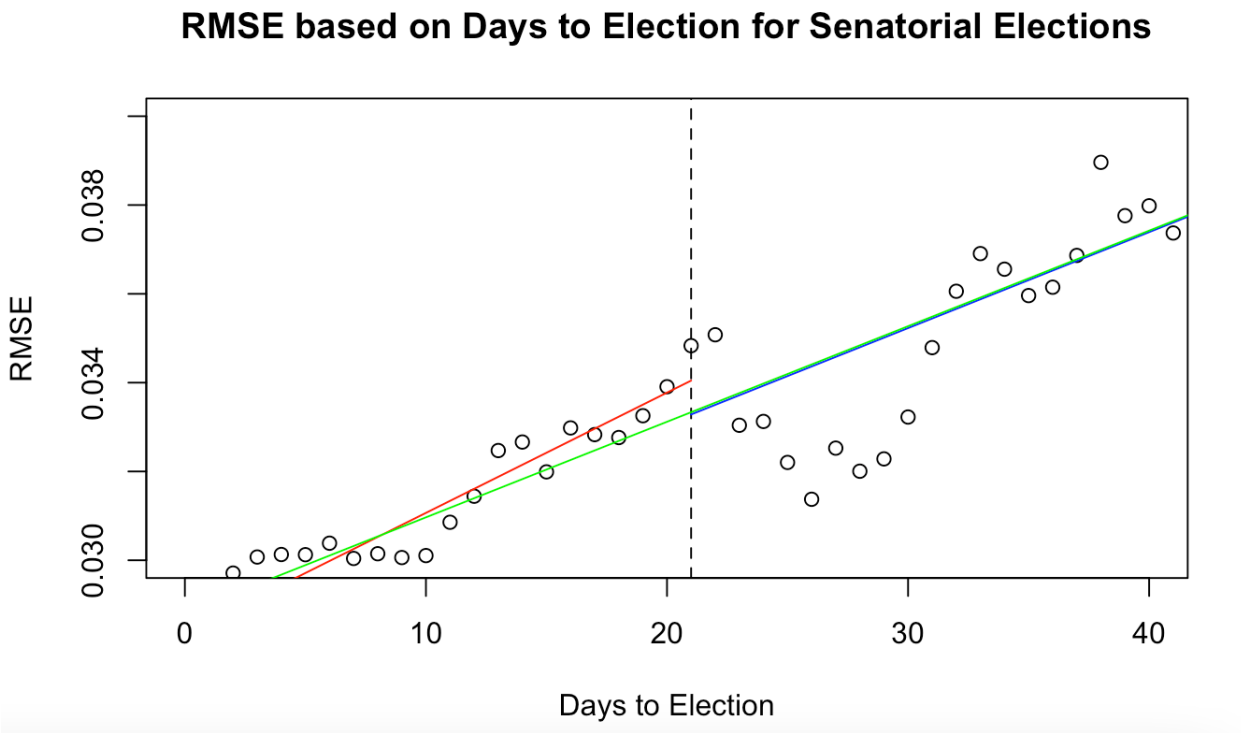
$$RMSE_{after} = \widehat{time}_{after} * X_1$$

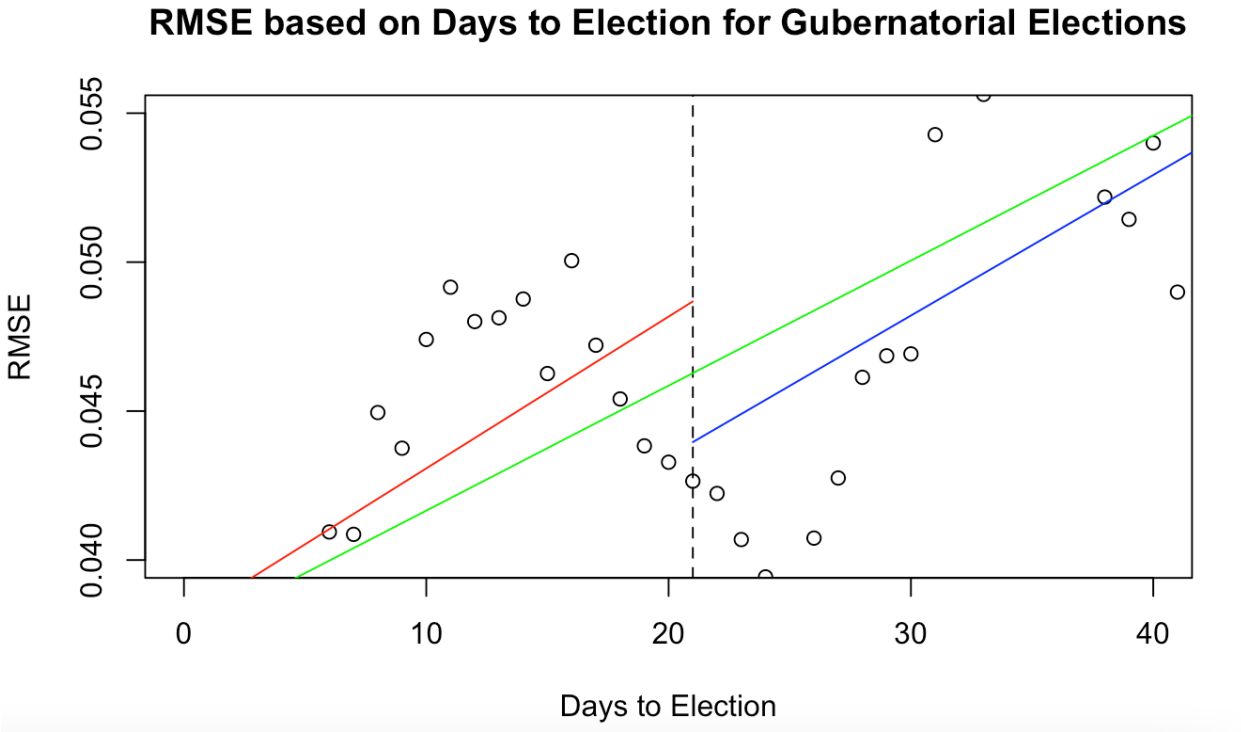
The set of hypotheses used for the test are as follows:

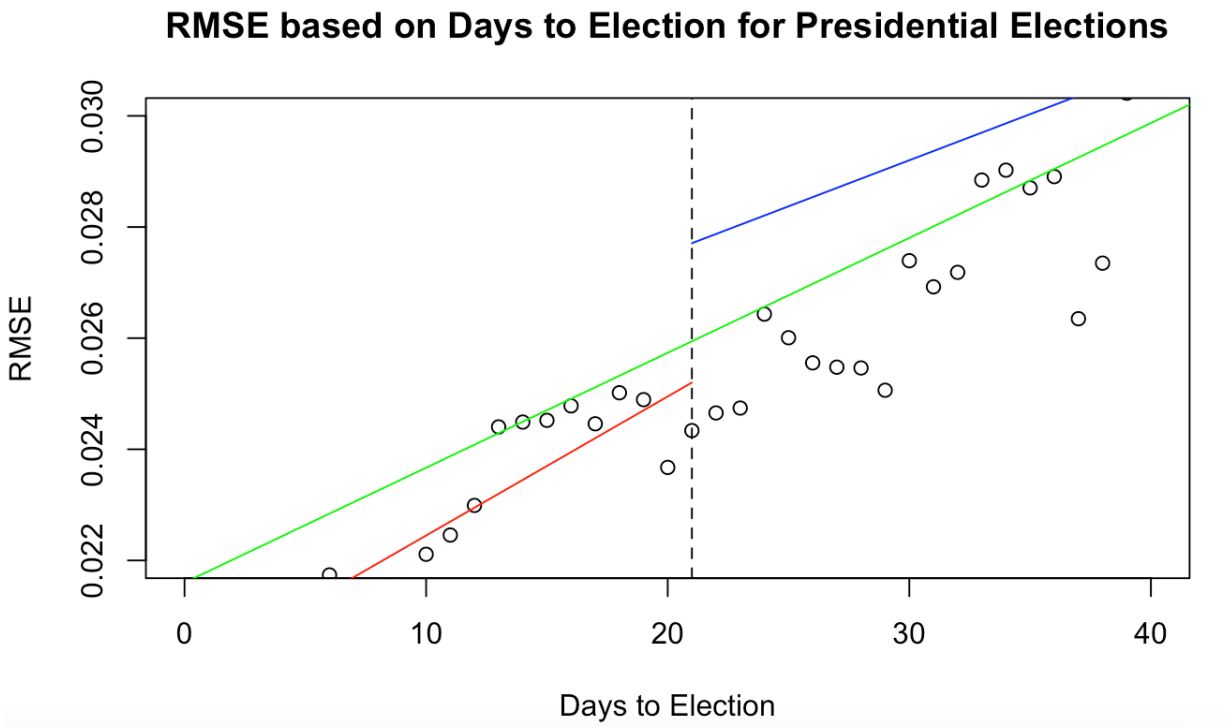
$$H_0: \widehat{time}_{before} - \widehat{time}_{after} = 0$$

$$H_A: \widehat{time}_{before} - \widehat{time}_{after} \neq 0$$

The auxiliary dataset was used to conduct the aforementioned linear regressions and Chow tests given its inclusion of polls dating 100 days prior to election day and for consistency with the original paper's plot of RMSE as a function of time. All three separate election-levels were first separated and individual Chow tests were conducted for each set to determine whether a statistically significant break existed for any of the three subcategories of polls. A third linear regression inclusive of all the data was conducted in each set of election-level polls for reference. The graphs below illustrate the two inferred true lines before and after the three-week mark:

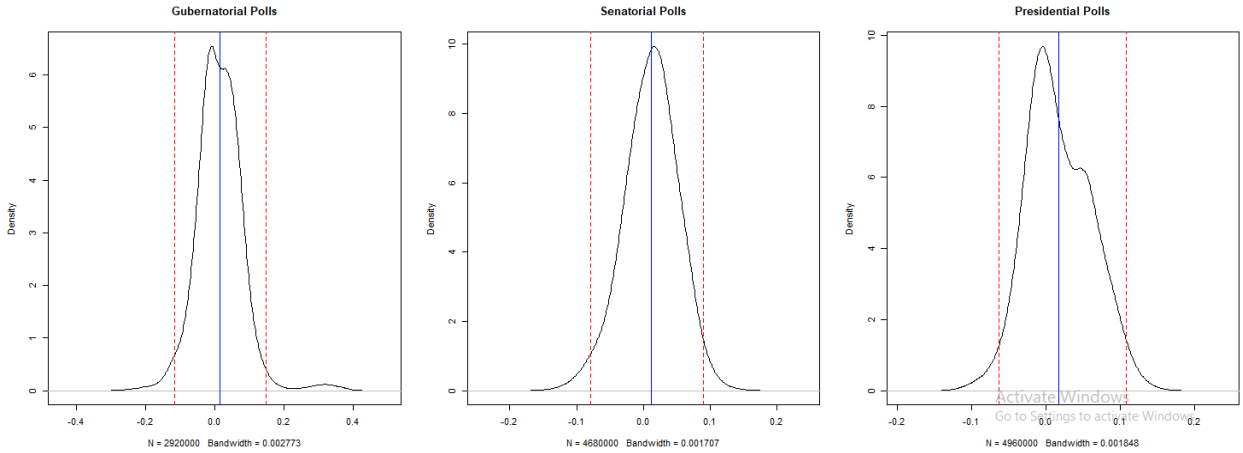

RMSE based on Days to Election for Senatorial Elections
RMSE
0.030
0.034
0.038
0
10
20
30
40
Days to Election


RMSE based on Days to Election for Gubernatorial Elections
RMSE
0.040
0.045
0.050
0.055
0
10
20
30
40
Days to Election



With regard to senatorial election polls, the two lines corresponding to before and after the three-week mark appear to align well with the green line representing the inferred true line inclusive of all data points. As such, there does not appear to be a major difference between the two which suggests that we would fail to reject the null hypothesis that the true coefficients of the two lines are the same. The Chow test for senatorial polls reflects similar results with the test producing an F-statistic of 0.12073 and a p-value of 0.8864. With regard to gubernatorial polls, the line corresponding to the rate of change in RMSE within the three-week window appears to actually have a steeper slope than the line after which seems to suggest that time-dependent bias may be more influential within the three-week period. However, overall, the lines do not appear to be significantly different from one another with the Chow test producing an F-statistic of 1.2487 and a p-value of 0.292. With regard to the presidential polls, the slope of the line corresponding to within the three-week mark seems to also have a slightly greater slope than that of the line corresponding to outside of the time window. However, the lines do not appear to be distinct enough to justify a statistically significant difference with the Chow test producing an F-statistic of 1.4869 and a p-value of 0.2318.

Given the fact that none of the Chow tests produced statistically significant results, such a conclusion suggested the three-week window may obscure from the potential effect of time-dependent bias. As such, in order to gauge the effect of time-dependent bias irrespective of the three-week window, the total number of β terms in each election-level from all elections and all iterations (number of elections in election-level * 40,000) was gathered and sorted to generate a 95% confidence interval. While credible intervals are generally used in Bayesian analysis, confidence intervals were used as a substitute for ease of calculation. After all β terms, were collected, density plots were constructed as shown below (the N in each plot is larger than stated before as we are using the auxiliary dataset inclusive of polls beyond the three-week window):

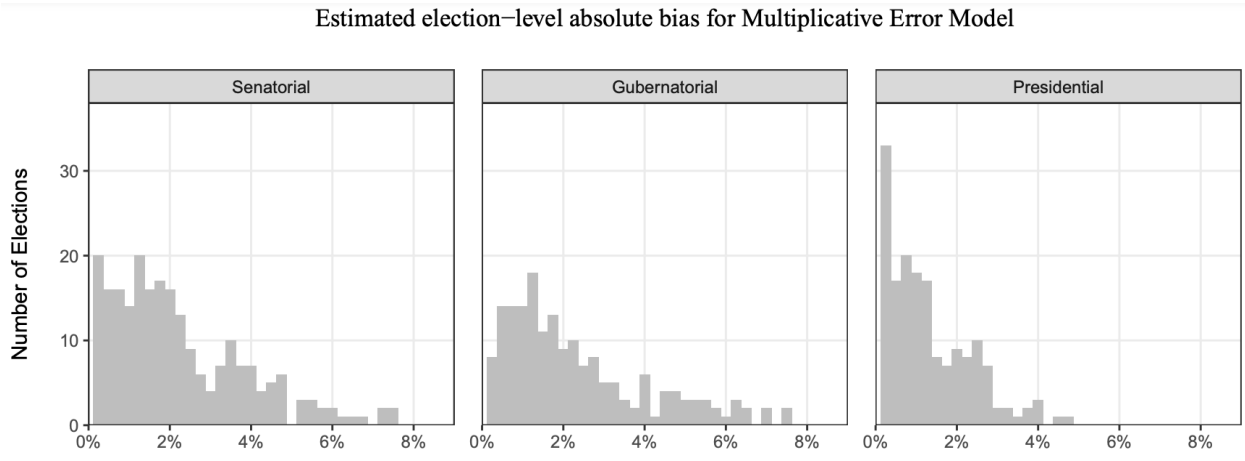


It is clear from first glance that all of confidence intervals, as delineated by the dashed red lines, contain 0.0 with the mean generally hovering around 0.0, as well. As such, it it can be assumed that the time-dependent component of bias is actually negligible even when considering polls beyond the stated three-week window. In terms of the model, such results imply that within the bias term ($\alpha_{r[i]} + \beta_{r[i]} t_i$) must be mostly composed of the fixed, systematic effect of $\alpha_{r[i]}$, and that removal of the time-dependent term should not qualitative change estimates of average election-level bias. Such a discrepancy may be due to the technical constraints of this report given the use of confidence intervals instead of credible intervals. However, such a discrepancy does not necessarily imply a contradiction. The statistical insignificance of the Chow tests suggests that the specific three-week window itself is arbitrary and does not constitute a definite hard cutoff. As such, the confidence intervals may actually be seen as an extension of the

implications of the Chow tests in that not only is the three-window arbitrary but also that any time-dependent constraints on the dataset within 100 days of election day can be dropped without significantly affecting estimates of average election-level bias. In either case, the results of the sensitivity analysis suggest that the original paper's assumptions regarding time-dependent bias, while arbitrary, are robust.

Multiplicative Excess Variance:

With regard to the additive excess variance assumption, the original paper had stated that the assumption was made due to ease of calculation and modelling and not on theoretical grounds. To test such an assumption, this report had changed the original model sampling variance from $\sigma_i^2 = \frac{p_i(1-p_i)}{n_i} + \tau_{r[i]}^2$ to $\sigma_i^2 = (\frac{p_i(1-p_i)}{n_i}) * \tau_{r[i]}^2$ and attempted to fit the new model on the primary dataset used in the original paper. Qualitatively, such a evaluation makes sense in light of standard methodological practices among polling firms with polling firms generally employing some form of post-sampling multiplicative weight adjustment when calculating point estimates of predicted vote-share in an election (not when calculating the SRS-based margins of error of such point estimates, though). The results of fitting the altered model onto the original dataset is visualized below, with each histogram datapoint on the histogram representing b_r:



The resulting estimates for average election-level absolute bias are as follows: 2.4% for senatorial election polls, 2.5% for gubernatorial election polls, and 1.4% for presidential election

polls. Comparing the estimates generated by the multiplicative model and the original model reveal similar trends with regard to increasing bias when observing lower-level elections. However, while the general trends remain roughly similar to results produced by the original model, the results generated from the multiplicative model reflect a higher estimated average election-level bias and greater spread in election-level absolute bias. While not shown in the histograms, the standard deviations of the estimated calculated average election-level absolute bias were calculated to be 0.064 for senatorial election polls, 0.067 for gubernatorial election polls, and 0.056 for presidential election polls. Comparatively, these are lower than the standard deviation of the posterior point estimates produced by the original model would imply greater confidence in the estimates generated by the multiplicative model. However, such interpretations of the difference in standard deviation may not hold given the fundamental differences in the models (i.e. distribution of σ_i^2 would be different which would change calculations of p_i, etc.).

To determine whether such an increase in average election-level bias for each election-level constituted a significant difference, the multiplicative model and the original model were fitted onto the original dataset and bootstrapped to proximate the difference in average bias in election *r[i]* (the draws themselves could be interpreted as samples, hence bootstrapping). T-tests were not feasible considering the non-frequentist interpretations of the mean and standard deviation in Bayesian inference. More specifically, given that each election has 40,000 corresponding sets of parameter values, a sample was drawn from the posterior distribution of estimated b_r values for each election (each of the 40,000 parameter sets generates a new p_i which is used to compute b_r). This process can be repeated twice for every election given that both models are fitted on the same data. As such, for each election in the chosen dataset, two random samples of estimated b_r values were drawn from the posterior distribution of b_r to constitute a set of arbitrary random pairs from which differences can be calculated. Given the random nature of the sampling, these calculated differences between the pairs was taken to be representative of the true difference among the two models for each given election. The 2.5% and 97.5% quantiles were then found to construct a 95% confidence interval for the difference in b_r for each election. If the 95% confidence interval contained 0, then it was concluded that there existed no significant difference between the two model's b_r terms. Such a

process was repeated for every election in the dataset. For the sake of computational expediency, only the presidential poll dataset was used.

With regard to the actual results, 99.44% of the confidence intervals/elections of the presidential elections in the primary dataset did not contain 0. Such a high proportion was most likely generated due to the high number of iterations when fitting the model, and the high sample size used when bootstrapping (40,000 to correspond with a complete shuffling of the posterior distributions of both models). While a specific hypothesis test was not used, such a high percentage of non-significant confidence intervals suggested that the two models perform qualitatively similar despite the multiplicative model producing a slightly greater estimate of average election-level bias. As such, the similarities between the two models suggests that the presented model is robust not only to alternations to the dataset, but to model variations as well. In terms of theory, such a result could suggest that while excess variance varies from polling firm to polling firm, differences may be generally negligible when aggregating hundreds of different polling results. Such a result may also have been due to the fact that presidential elections were used as the basis of comparison given that the original polls had shown that presidential polls exhibited less variability in election-level absolute bias compared to senatorial and gubernatorial polls given the significant resources dedicated presidential polls.

Conclusion:

Overall, this report sought to reproduce the results from *Disentangling Bias and Variance in Election polls* and sought to the evaluate the results both in terms of the dataset used to train the model and the structure of the model itself. The original results were able to be reproduced with a high degree of accuracy with the original measures of average election-level absolute bias for each election-level being 2.1% for senatorial polls, 2.3% for gubernatorial polls, and 1.2% for presidential polls which aligned exactly with the original results. Such congruity can be attributed to high number of iterations when fitting the model onto the data (set.seed was also used but was only relevant in reproducing the SRS-based polls used in this report's data exploration. Subsequently, sensitivity analyses were carried out to evaluate if whether the results were robust subject to changes in the dataset used to fit the model and the structure of the model itself. When changing the dataset to include polls dating within 100 days of election day instead of the assumed three-week window, the estimated 95% confidence interval for β was found to

include 0 which ultimately reinforced the authors' assumption that time-dependent bias may be negligible despite the three-week window itself being arbitrary. Changing the additive excess variance term yielded average election-level absolute bias estimates higher than those of the original paper but such differences were found to be negligible after bootstrapping the posterior distributions of b_r terms for each election for both the original model and the evaluative model with multiplicative excess variance. While such evaluations reinforce the validity of the findings and model presented by the authors, future research may continue to test some of the fundamental assumptions made while also expanding their findings onto a wider range of polls such as House election polls and non-US elections polls.