Capstone Final Report: Microsoft FY1830, Team 1

Kathy Lin, Jie Lu, Tin Oreskovic, Jake Snyder

12/16/2018

Abstract

Polling is one of the most robust methods to capture and understand public opinion, especially in politics. However, as with all sampling methods, there are limitations and biases, particularly with regard to the questions asked. We investigate a new question format for political opinion polls with the goal of improving the effectiveness and predictive power of these polls. Using a mobile polling strategy, we ask respondents demographic information, as well as who they plan to vote for in the 2018 midterm elections. Our research uses multilevel regression and poststratification to control for the well-known biases within polling data, such as sampling bias and nonresponse bias. We use Bayesian methods to predict two-party share of the 2018 midterm Congressional vote. Our research shows that, with respect to predictive power, the new question format does not differ much from the traditional format. Since the new question format requires more time and money than the traditional format to implement, it does not seem worth the investment.

Introduction

The 2018 U.S. Midterm Elections took place on November 6th, 2018 and, beyond determining the balance of political power in and the makeup of the House of Representatives and the Senate, served as a pulse check for any medium and long-run potential changes in the political landscape of the United States. As a result, the Democratic Party won control of the House of Representatives; and with it, the ability to check President Trump's power. Republicans, however, expanded their majority in the Senate by performing better among the limited number of Senate seats available on the ballot for these Midterm Elections.

We decided to investigate whether when using non-representative online polls to predict the vote share for the majority party, specifying candidate names in questions leads to higher predictive accuracy than when generic ballots are used. That is, is there a substantial change in the ability to predict the majority party vote share between a generic poll that asks whether a participant will vote Democrat or Republican (generic ballot) and a poll asking a participant whether they will vote for, say, Democrat Harley Rouda or Republican Dan Rohrabacher (specific candidate-named ballot)?

Nationwide surveys have been widely used for measuring public opinion, evaluating public policies and, ultimately, predicting elections. There are, however, various risks associated with traditional surveying methodologies that limit the interpretation of results from election forecasts. If these risks are not mitigated, the results from the surveys can mislead the public and further elevate the existing levels of public mistrust in surveys (caused in part by the disparity in the public's impression of the uncertainty of the average projections and the actual level of uncertainty of the projections of the 2016 Presidential Election). The benefits of using online survey data, on the other hand, are their cost-effectiveness and time efficiency; an opt-in online survey requires less than one-tenth the time and money used to reach the same volume of participants in a traditional survey. However, these online surveys suffer from a variety of sampling and non-sampling errors, such as frame, nonresponse, measurement and specification errors - some of which are not mirrored in the issues of traditional, probability-based polls. At the same time, the advantages to forecasting elections with representative polls are diminishing due to falling response rates (Wang, Rothschild, Goel, Gelman; 2015), which makes using online polls yet more attractive.

To mitigate the biases introduced by online, non-representative polling and polling more broadly, we have turned to a state-of-the-art methodology first developed by Gelman and Little (1997) for national/subnational survey data analysis - multilevel regression and poststratification (MRP).

Our focus, then, is to first (i) "directly" predict the two-party vote shares across the districts, using multilevel regression and poststratification separately on candidate-specific and generic data, and (ii) compare the two models' performance to investigate whether there is a performance-benefit to including the more expensive candidate-specific questions into polls.

Related Work, Contributions

Election polling, including online survey results, are subject to a variety of sampling and nonsampling errors, as mentioned earlier. It is also naive to attribute the discrepancies between poll results and actual election outcomes purely to sample variance. According to past work by Biemer, Groves and Lyberg in 2010, there are at least four additional types of errors that aren't reflected in margins of error: frame, nonresponse, measurement, and specification errors. In Disentangling Bias and Variance in Election Polls (Shirani-Mehr, Rothschild, Goel, Gelman), the errors are defined as follows. Frame error occurs when there is a mismatch between the characteristics of the sample and the target population of interest. For instance, in election surveys, if the sampling frame can include many people who aren't likely to vote, this needs to be corrected with likely voter screens, which are typically estimated with error from survey questions. Nonresponse errors are a growing concern and occur when missing values are not missing at random but rather systematically, related to the response. As an example, supporters of a trailing candidate may be less inclined to respond to surveys (Gelman, Goel, Rivers and Rothschild). Measurement error arises when the wording of a survey question or the ordering of the questions affect the response rates (Smith, 1987). Lastly, specification error occurs when a survey respondent misinterprets a question from what the surveyor intended to convey.

Within the past 100 years, there have been many attempts to forecast the outcome of elections based on public opinion polls, with probability-based, representative polling seen among US mainstream media as the only legitimate form of polling since the mid-50s. The representative, probability-based surveys during the final week before an election have done reasonably well in forecasting the outcome of the popular vote using Root-Mean-Squared Error as the loss metric, between 1936 and 2009 the forecast was only 2.72 points off from the true outcome (Erikson and Wlezien, 2012a). When aggregating representative surveys to predict the final outcomes of elections in the US (rather than the raw popular vote outcome) overall and by states, however, forecasters face a host of obstacles: most prominent among these is a reliance on estimates of polling data with no access to underlying individual-level survey responses, which can lead to substantial misses in contiguous state predictions with correlated errors. (Konitzer, Corbett-Davies, Rothschild; 2016; Non-Representative Surveys: Modes, Dynamics, Party, and Likely Voter Space)(http://www.realclearpolitics.com/articles/2016/11/12/it_wasnt_the_polls_that_missed_it_was_the_pundits_132333.html).

More recently, a body of literature emerged on using non-representative surveys, much like the one used in this project, to estimate public opinion, as well as to forecast election outcomes. In each of these, the relatively much lower cost and greater speed of collecting online opt-in samples is emphasized as the primary pragmatic advantage over traditional survey methods. In Non-Representative Surveys: Fast, Cheap, and Mostly Accurate (Goel, Obeng, and Rothschild; 2014), an opt-in sample on 59 attitudinal questions is collected, and, after model-based poststratification adjustments, the estimates are compared with estimates for equivalent questions from the 2012 in-person General Social Survey and phone surveys by the Pew Research Center the median absolute difference of 7.4 percentage points between the estimates from the first (online) and second (GSS) estimates is comparable to that between the GSS estimate and the Pew estimates.

Given the diminishing advantages of collecting probability-based, representative polls due to many of the above-mentioned errors, especially due to non-response, forecasting elections with non-representative online polls seems an increasingly attractive option (Wang, Rothschild, Goel, Gelman; 2015). Adjusted through multilevel regression and poststratification, the 2012 presidential election forecast by Wang, Rothschild, Goel, and Gelman yields accuracy comparable to that of leading poll analysts using traditional methods, which are based on aggregating hundreds of representative polls. Non-Representative Surveys: Modes, Dynamics, Party, and Likely Voter Space (Konitzer, Corbett-Davies, Rothschild; 2016) looks to address the above-discussed correlated error between the state-forecasts for the 2016 presidential election as a result of aggregating traditional surveys. Using a non-representative mobile-only single panel poll, the authors employ a dynamic multilevel regression and poststratification approach, which can disentangle changes in sample composition over time from genuine changes in responses - an important distinction since the former are often misinterpreted as the latter due to overinterpretation of recent polls.

As the other sections herein specify, we adopt many of the elements of this approach that thus far produced comparatively accurate forecasts using non-representative samples, namely of first modelling the probability that a respondent will vote for a candidate of either of the major party candidates (the question may refer specifically to the candidate or only the party), followed by a projection of these estimated probabilities onto the available data on the demographic composition of the population in each of the districts from which we have polling responses. Most simply, our findings add to the above-reviewed literature by replicating desirable forecasting results using the methods previously found to be effective on a new, politically highly relevant dataset. Beyond the benefits of attempting to replicate results of known methods, applying the methods to a slightly different domain, specifically to Congressional elections, is another contribution to extending the literature.

More distinctively, however, and more importantly for the aim of our project, our direct contribution is the comparison of asking respondents questions that refer to specific candidates to asking only about the major parties. Given the higher cost of incorporating specific candidates' names into the polls across the districts, we investigate whether there is any benefit to doing so in terms of prediction accuracy. If the improved accuracy is large enough, this would potentially justify the higher costs. As the results section explores in greater depth, according to our findings there is no benefit to doing so with respect to general forecasting performance.

Discussion of the Data Set and Exploratory Data Analysis

With the complete collection of polling surveys, our final datasets consist of three parts:

1. Generic House Ballots: ranging from November 2017 to October 2018, which has over 200000 observations. Note less than half of these observations have district-specific data which is needed for our comparative analysis between generic and specific house ballots' relative predictive power for two-party vote share.

2. Specific House Ballots: The number of responses in each of the 40 specific ballots varies widely, from 200-400 each. The total number of observations is over 9000.

After excluding "third option" and "undecided" responses and downsampling the generic dataset (to allow performance comparisons, as explained more elaborately below) both datasets have 6209 observations.

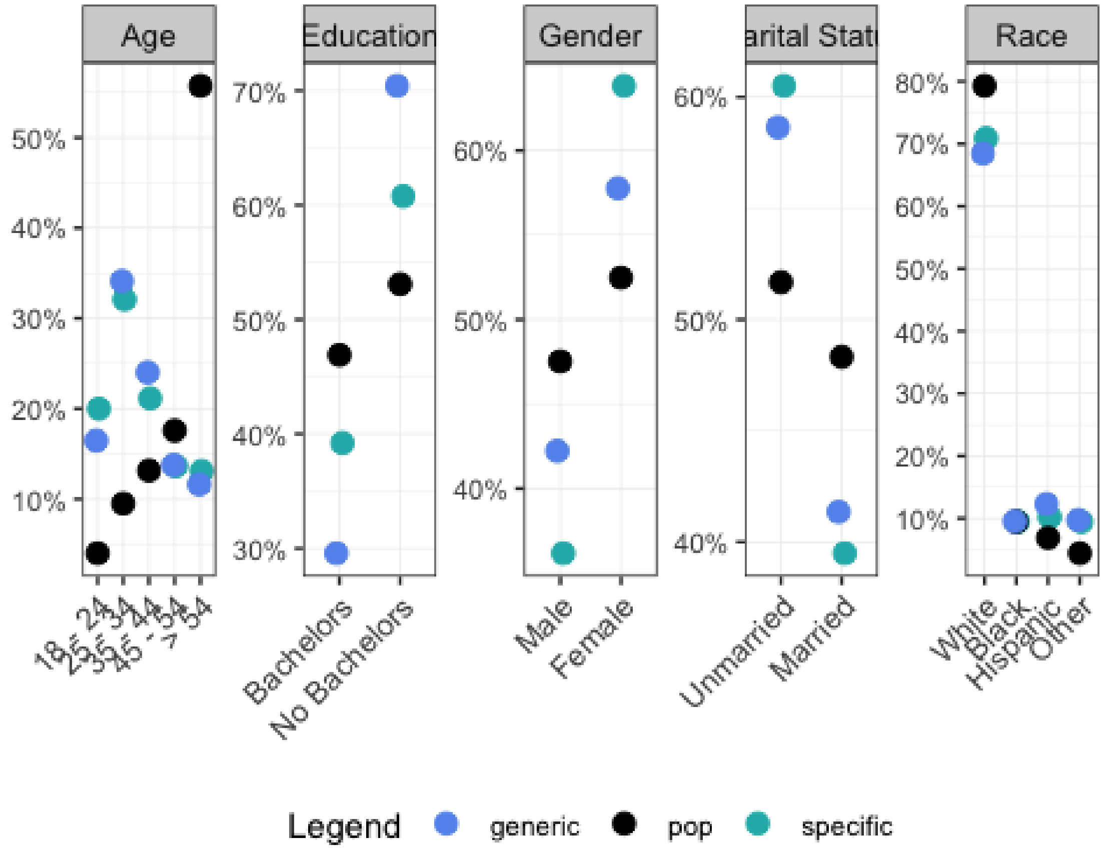
3. Poststratification space: dataset was prebuilt and preprocessed across several demographic variables such as age, gender, race, education, marital status, geography (state, district, urbanicity), and party affiliation. It estimates the space of likely voters for the 2018 US House of Representatives Elections.

Both the generic and specific datasets investigate respondents' attitudes, demographic and geographic information. Demographic features such as age, gender, race, education level, income level, and geographic location are used to break the population down into hierarchical levels in our analysis for a level-wise vote turnout rate forecasting. The generic ballots survey consisted of 14 demographic questions and 3 attitudinal questions, while the specific ballots survey consisted of 14 demographic questions and 17 attitudinal questions, which aim to dig deeper into respondents political attitudes towards a specific candidate of a major party. The most significant difference between generic and specific ballots data is the usage of specific candidate names. The specific candidate ballots values have also been preprocessed (specific names are omitted by mapping candidate names back to their respective political party) by our mentors so that they may be row-binded to the same dataframe.

The column of interest from the specific ballot includes: “If the election for the U.S. House of Representatives in your district was today, who would you vote for?”

As mentioned earlier, participants in these surveys are self-selecting. Thus, the first part of our review of the newly added Pollfish polling data (generic ballots) and specific ballots consisted of exploratory data analysis and visualization focused on issues emerging from potential imbalances present in the datasets. Each of the graphs below are further segmented by the vote2018 column in the dataset, which corresponds to “If you were to vote in the US Midterm elections 2018, who would you vote for?”. We note that the specific polling has mapped specific candidate names back to their respective parties.

Figure 1



Proportion of Respondents by Demographic Variables for Each Dataset

We begun our exploratory data analysis by examining the potential data imbalances across our three datasets: population, generic polling data and specific polling data. The above plot exemplifies the aforementioned data imbalances as well as the difference in imbalances across datasets. The graph is facetted on different demographic variables, where each facet’s categories (for example: Gender has two categories – Male vs Female) plots the corresponding dataset’s proportion of respondents in that category. The light blue dot represents the proportion of respondents in the generic polling dataset. The green dot represents proportion of specific polling respondents. And finally, the black represents the population data’s proportion. For example, every dataset is skewed the same way across the following demographic variables: Education, Gender, Marital Status and Race. It is clear that racially, our polling data and the population data is heavily dominated by Unmarried, Female, White voters with no Bachelors degree. It’s interesting to note that the population dataset, however, differs greatly from both of our polling datasets in that the >54 age group is the dominant age group for respondents. For the specific data, due to limited number of respondents and different survey techniques, features are far away from the true population; for the generic data, even though it’s close to the real distribution for some features, when it comes to education and gender, the dot still deviates from the true population.

Given that the trends in the generic polls and specific polls were similar across demographic variables, we wanted to explore any potential differences geographically between the generic and specific. More specifically, we explored the proportion of respondents of each survey type who are likely to vote Republican. The results are shown below.

Figure 2

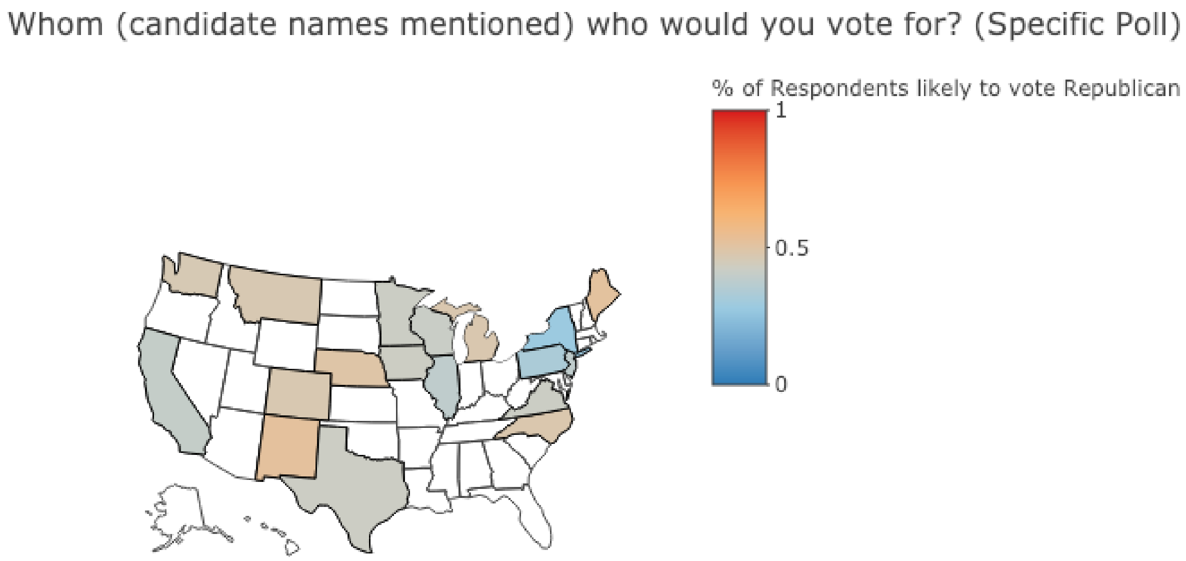
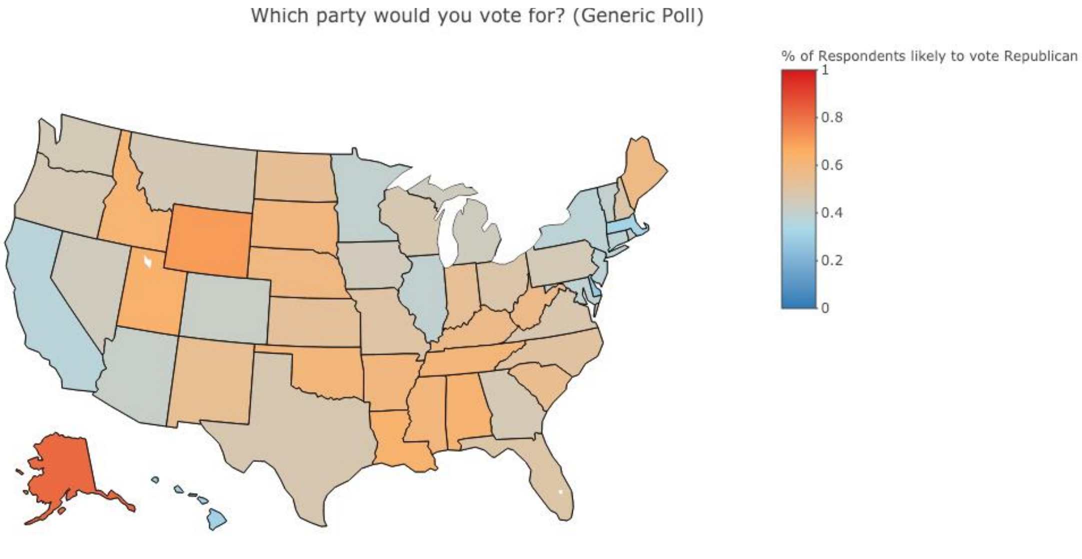

Whom (candidate names mentioned) who would you vote for? (Specific Poll)
% of Respondents likely to vote Republican
1
0.5
0

Figure 3


Which party would you vote for? (Generic Poll)
% of Respondents likely to vote Republican
1
0.8
0.6
0.4
0.2
0

Note that although our specific data was collected from 40 districts that were likely to lean towards the majority party, as shown in the specific map above, the proportion of respondents likely to vote for the Republican in the 2018 midterm elections surprisingly were mainly less 50%. It is also clear that generic polling resulted in a more diverse pool of responses due to the greater number of respondents we have for generic surveys.

With respect to respondents' Age, we've found that both specific and generic polls showed a majority of respondents would likely vote for the Democratic candidate; However, in generic poll's >54 age group, the proportion of voters likely to vote Republican was greater than that likely to vote Democratic (opposite in trend to the specific polling results). With respect to respondents' Education, it's interesting to note that a greater proportion of specific poll respondents had Bachelor's degrees (regardless of their political affiliation) compared to generic polling respondents. Although both generic and polling datasets' respondents mainly identified as White, the political affiliation of the two datasets' differed tremendously; White generic polling respondents were more likely to vote Republican while white specific polling respondents were more likely to vote Democrat. Lastly, a greater proportion of married respondents in generic polling were more likely to vote Republican than married respondents of the specific polls.

Multilevel Regression and Poststratification

As specified in the introduction section, we looked to answer two questions, the first of these in order to allow answering the latter:

Is there an efficient and effective way of fully utilizing online poll data to predict two-party vote share in Congressional elections? Is polling data with specific candidate names listed worth investing time/money to collect?

It's naive to believe that nationwide surveys can achieve full population coverage. Therefore, we've chosen to mitigate the nonresponse bias risk by modeling the respondents' responses and using poststratification; specifically, employing the multilevel logistic regression and poststratification (MRP) methodology. At a high level, poststratification involves modifying the weights based on population counts such that the sample is recalibrated to the true population counts in the poststratification (Little, 1993). This method relies on the representative population data taken from voter files and the census. MRP has been used to recover state-level estimates of public opinion by projecting the output of a multilevel regression model onto the space of likely voters. This method has a history of over 50 years history (Pool, Abelson and Popkin, 1965), and is adjusted to fit nowadays' survey techniques. (Park, Gelman and Bafumi, 2004). We begin by running multilevel logistic regression models on the poll samples, with its features matching demographic or geographic predictors. Subsequently, the poststratification step weighs the estimates for each demographic or geographic respondent category according to the number of people in fact fitting this category in the selected level according to the available population data.

The models meant to predict the respondents' voter-intent that we design make use of the the following independent variables: age, gender, race, education, urbanicity, marital status, and political affiliation. The selected level in our case is be the congressional district code, ranging from AZ00 to WY00. The population level data is obtained from the project advisors, and the projection space for US national voters is carefully cleaned and parsed by PredictWise, ready to represent the population of 2018 voters.

There are several benefits to using a multilevel model (here a multilevel logistic regression). One of these are the so called partially-pooled estimates of the parameters: a separate, no-pooling model (expressed by adding variables and interactions) for each state and district would likely lead to overfitting the data due to a possible relative lack of data in certain districts, and would overstate the variation between them, while a classical, single-level regression (complete pooling) does not exploit the hierarchical structure of the data - voters really vote within district but differ in how their “features” correlate with their voting intentions across districts. In multilevel models, the partially pooled estimate can be close to complete pooling for groups with a small sample size and close to no pooling for groups with large sample size, performing well for both sorts of groups and improving the accuracy of the average estimate. Another important benefit of multilevel modeling is that because the group-specific estimates are shifted closer to each other, the models are more robust to the classical problem of multiple comparisons: comparing the standard errors to their specific estimates is more justifiable despite the large number of comparisons for specific variables across the groups (states and districts). Finally, one of the many remaining benefits of multilevel models is that they do not require dropping one of the binary variables indicating department-belonging with all the estimates having to be interpreted in reference to the omitted group; rather, again due to the benefits of partial pooling, all the districts/states can be kept explicit in the model, hence aiding interpretability without suffering from multicollinearity.

Based on performance comparisons of various model specifications as indicated by the deviance information criterion (DIC), we opted for the following design:

$$Pr(\widehat{republican}_i = 1) = \mathrm{logit}^{-1}(\alpha_{j[i]} + \beta_{j[i]} age_i + \gamma_{j[i]} gender_i + \delta urbanicity_i + \kappa_i race_i + \eta_i party_i + \theta_i education_i + \tau_i married_i + \epsilon_i)$$

The dependent variable indicates whether the respondent i in district j replied to the survey saying they will vote for the Republican candidate. The variables whose slope coefficients above have both i and j indices (age and gender) are those whose estimates, along with the estimate for the intercept, were allowed to vary across the districts. This specification was selected as producing the best trade off between computational complexity and accuracy, while having a theoretical foundation in assuming that how voting decisions are correlated with gender and age varies across districts (as well as the intercept).

The poststratification step was performed using simple weighted averaging, for all demographic subsets. The poststratification estimate of voter intent $\widehat{republican}_s^{PS}$ for each subpopulation level s, then, is:

$$\widehat{republican}_s^{PS} = \frac{\sum_{j \in J_s}^{J} N_j \quad \widehat{republican}_j}{\sum_{j \in J_s}^{J} N_j}$$

Here J_s is the set of all cells (subsets) that make a level s (say, a district). The left-hand side above stands for the vote share in cell j, and N_j is the size of the jth cell in the population.

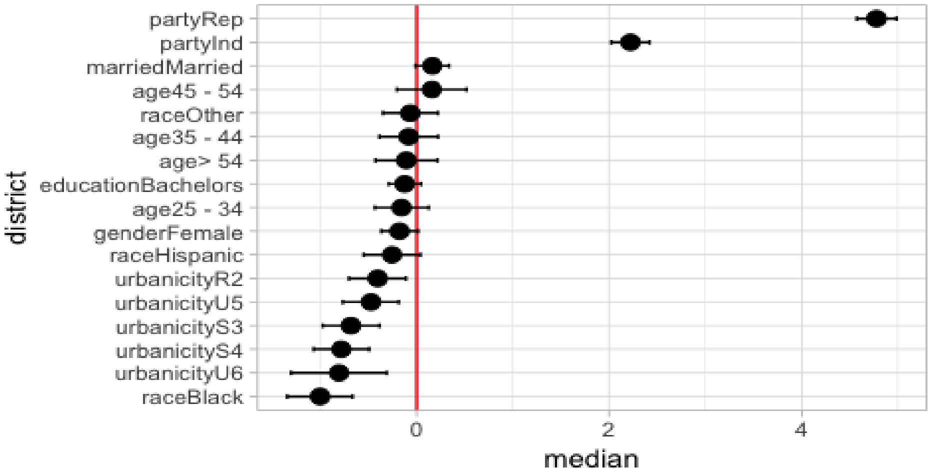
Ultimately, our aim was not to obtain the optimal predictions, but to compare the benefits of running either of the equivalent models on either the two datasets (one containing responses to candidate-specific questions, one to generic ballot questions), if there indeed is any difference. In order to allow a comparison between models as close to equivalent as possible, beyond using the same model specifications, we had to make a departure from what would produce yet more accurate predictions: we took a random subsample from the dataset with responses to generic ballot questions so as to exactly match the size of the candidate-specific dataset, each totalling 6209 observations once responses indicating neither of the major parties are excluded. Choosing a subset was meant to "level the playing field" for the two models, as well as to reduce computational costs while training on the generic poll dataset. Barring this adjustment and a possible trade-off between computational complexity and accuracy, we sought to optimize the MRP process for prediction accuracy, in order to allow a comparison of two models each as close to the optimal one given the data as possible.

Results and Discussion

Instead of examining directly the models' summary, we used the REsim() and FEsim() function of the merTools package to approximate/simulate the posterior distributions of both the non-varying and the varying parameters' coefficients.

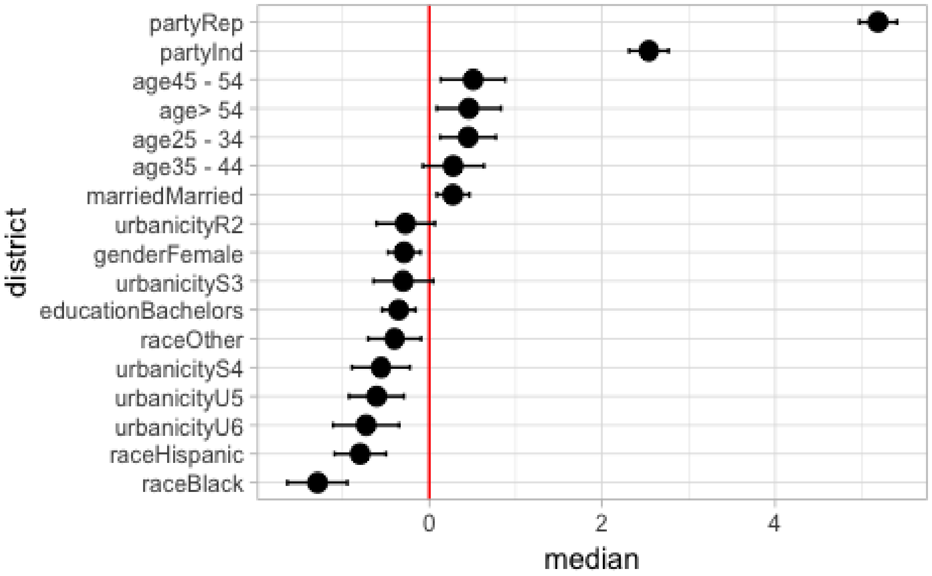
Approximated posterior simulations of the non-varying coefficient estimates from the specific model, visualizing how the various indicators correlate with an intention to vote Republican. The plots allow inspection of the size as well as the associated uncertainty for each estimate:

Figure 4



Approximated posterior simulations of the non-varying coefficient estimates from the specific model

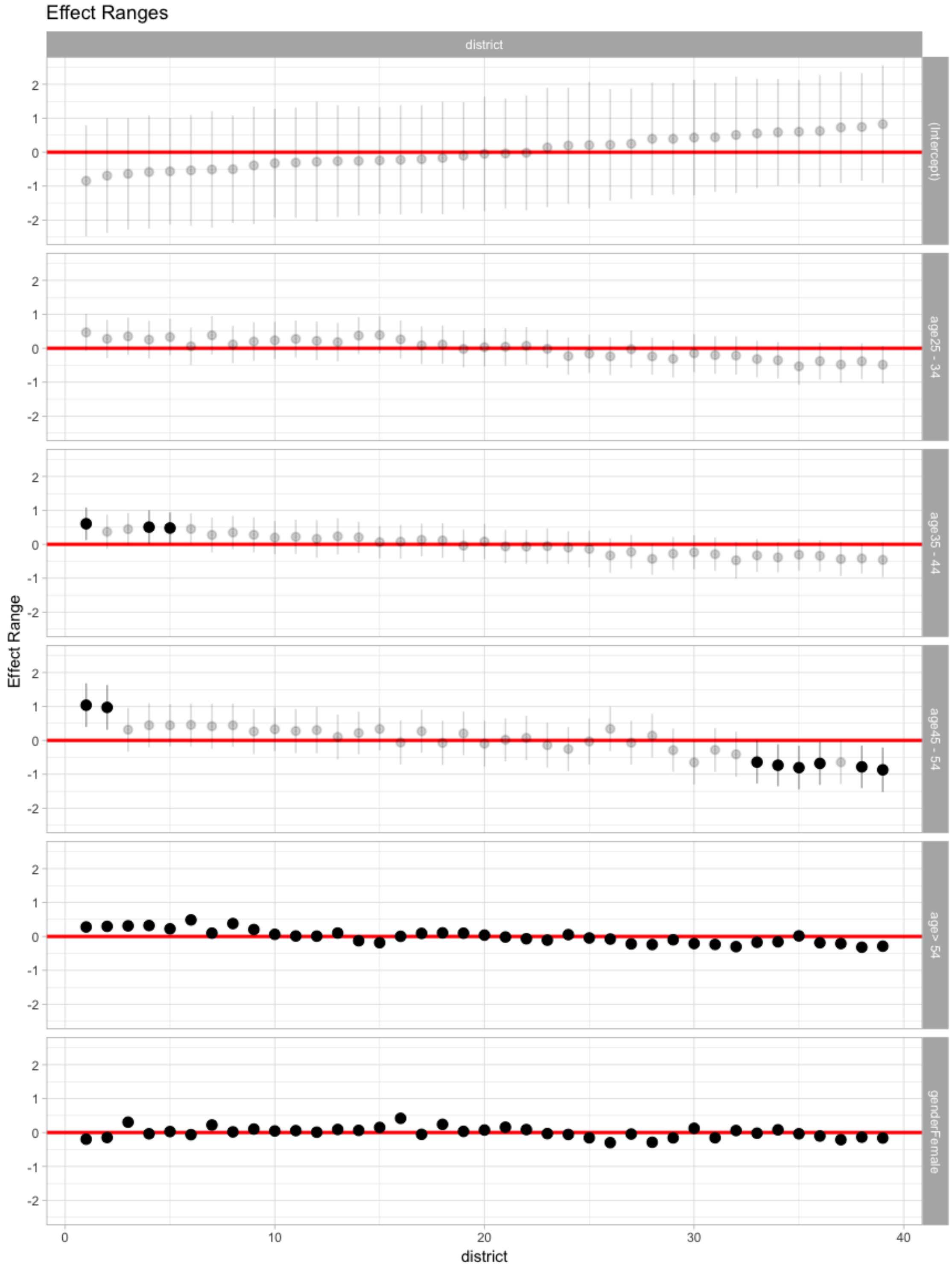
Figure 5



Approximated posterior simulations of the non-varying coefficient estimates from the generic model

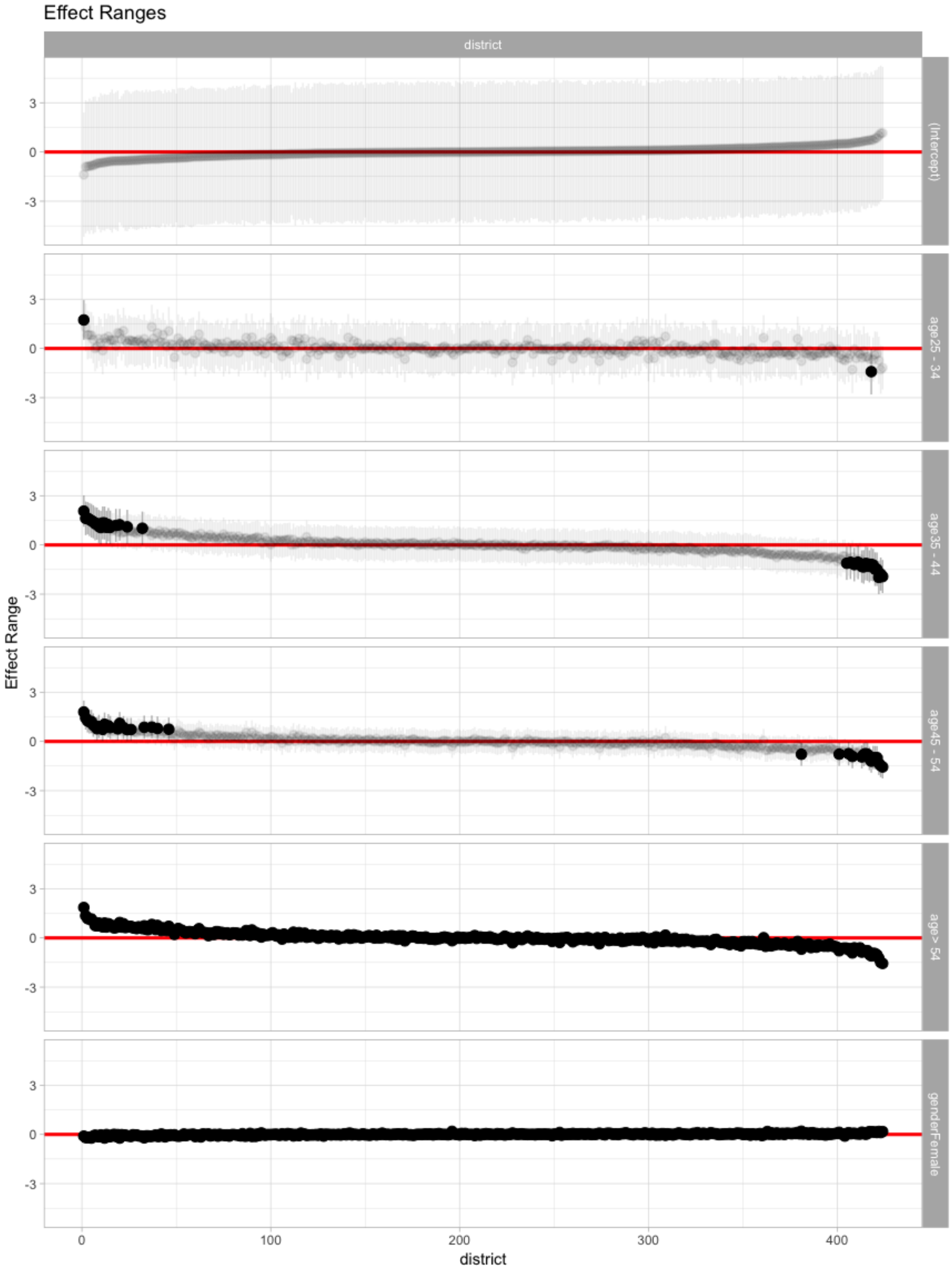
The two plots below show the approximated posterior simulations of the varying coefficient estimates, for the parameters that vary across districts. Estimates distinguishable from zero based on the inferred confidence band are highlighted. The variation shows what the multilevel model performs across districts, and it is clear that, since the coefficients in fact visibly vary across the districts, choosing a model that accommodates a hierarchical structure does influence the ultimate individual-level and (after poststratification) the district-level estimates of voter intent.

Figure 6



Approximated Posterior Simulations of the Varying Coefficient Estimates for the Candidate-Specific Model

Figure 7



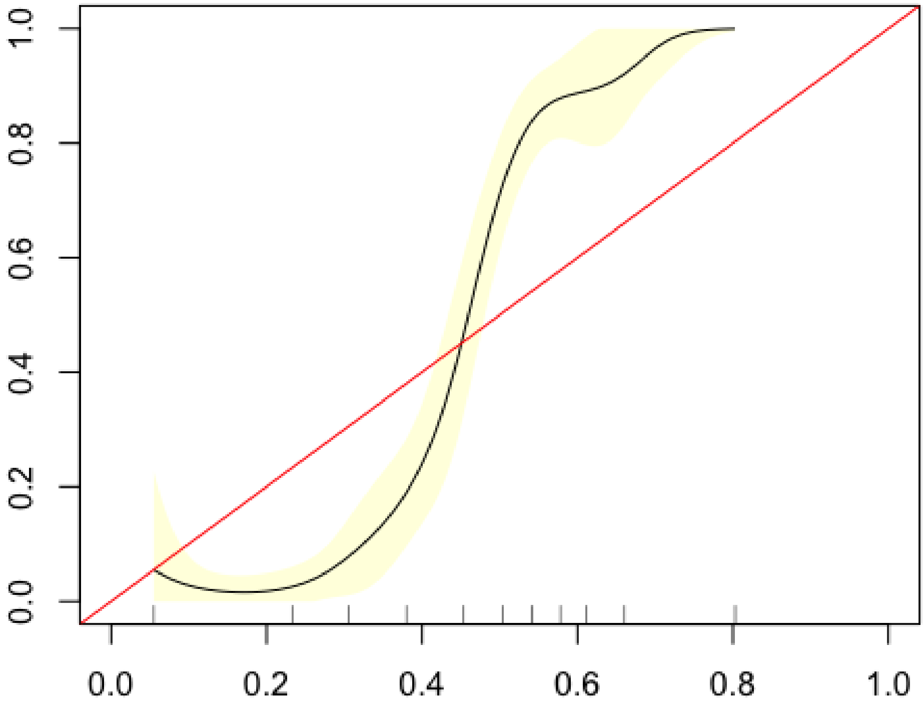
Approximated Posterior Simulations of the Varying Coefficient Estimates for the Generic Ballot Model

Model Comparison

Model	Data	Accuracy	RMSE	Precision	Recall	f1
Selected 40	generic	0.7297297	0.0704199	0.6428571	0.6428571	0.6428571
Selected 40	specific	0.6216216	0.0697910	0.4285714	0.5000000	0.4615385
All	generic	0.8891753	0.0674853	0.9095745	0.8680203	0.8883117
All	specific	0.8711340	0.0677441	0.8617021	0.8709677	0.8663102

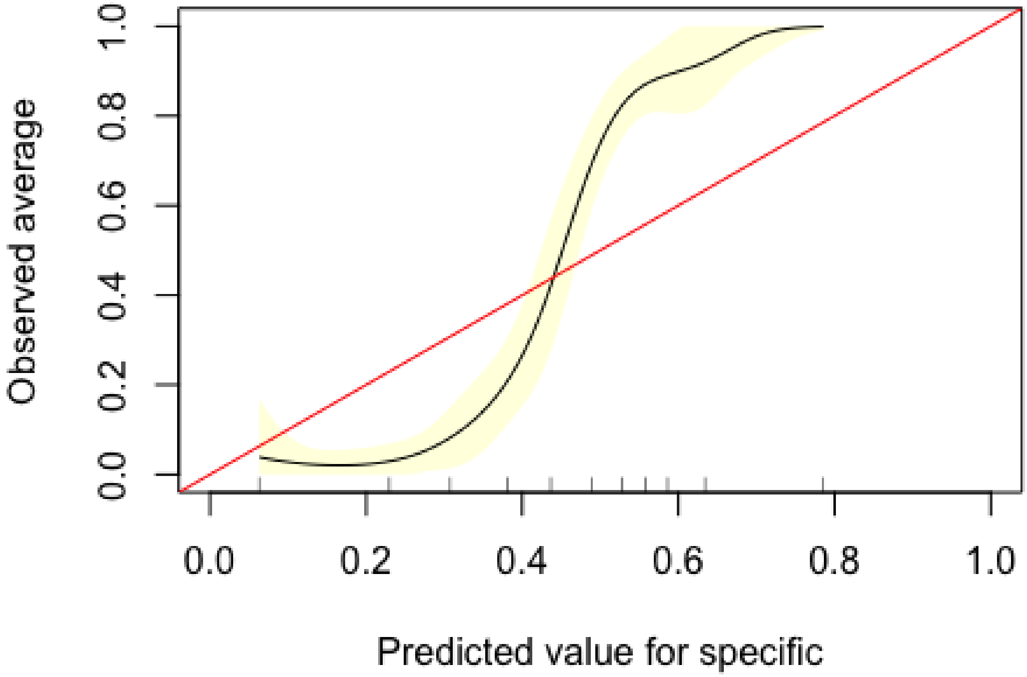
The table above shows the performance difference for the two datasets. We merged all the data available, processed it and treated it as the input for our MRP process. Using the polling responses to questions with specific candidate names, the MRP result would have a relatively better performance regarding accuracy and RMSE. We also examined the selected 40 districts where the specific surveys are distributed, and the result turns out to be similar to the one of all districts. The precision score is also interesting here. We can see these two input setting would lead to different precision score, which says among all respondents willing to vote for republican, generic survey would lead to a higher rate of true votes. And we infer that is the main reason why generic data could perform better than the specific data.

Figure 8



Calibration Curve for Generic Polling Dataset

Figure 9



Calibration Curve for Specific Polling Dataset

The plot above shows the calibration curve of our 2 survey results. A $y = x$ line is added for reference. In our settings with 10 bins, for each bin, the y-value is the proportion of true outcomes, and x-value is the predicted probability. A probabilistic model is calibrated if when binning the ground truth based on their predicted probabilities, each bin's true outcomes has a proportion close to the probabilities in the bin. In theory, a well-calibrated model has a calibration curve that hugs the diagonal straight line $y = x$. That is to say, a well calibrated classifier should classify the samples such that among the samples to which it gave a predict_proba value close to alpha, approximately alpha percent actually belong to the positive class.

Therefore, the model is off calibrated, especially when the predicted probability is high. We can see these two curves are in the shape of a typical transposed-sigmoid curve. This indicates that our model is over-confident. This may be because of the redundant features in our multilevel model. The square combinations of features has the potential threat to violate the assumption of feature-independence and result in over confidence.

The estimates using generic ballot polling questions perform marginally better than the estimates using specific candidate names. Since it is more expensive and time consuming to use specific candidate names in polling, we recommend continuing to use generic polling questions, as the investment in specific-candidate names does not seem to be worthwhile. However, further analysis is needed to determine whether this result holds within specific sub-demographics, such as age or education brackets.

Conclusions, Limitations, and Future Work

The estimates using generic ballot polling questions perform marginally better than the estimates using specific candidate names. A well-designed generic ballot survey can be applied to all congressional districts without any additional cost, however, a survey with specific candidate names will have to be adjusted before applying to every district, which would increase the overall cost and slow down the collection. Since it is more expensive and time consuming to use specific candidate names in polling, we recommend continuing to use generic polling questions, as the investment in specific-candidate names does not seem to be worthwhile.

However, further analysis is needed to determine whether this result holds within specific sub-demographics, such as age or education brackets. Besides, this study is mainly focusing on predicting the voter share with public polling data. Polling data, no matter generic or specific, may serve a wide range of other purposes. Especially, specific regional surveys would be beneficial to help address other regional issues. A long lists of interesting questions are asked and not covered for the purpose of this study. For example, "Is your community better off today", "Is today a good time to run a business in your community" and so on are all meaningful questions which make specific surveys valuable. Therefore, if the only purpose of the polling data is to predict the voter share, we would recommend generic polling because of its lower price and higher efficiency. If the surveys are designed to serve for multiple purposes, a further study would have to be conducted to estimate their overall pros and cons.

We limited the data we included in our datasets in order to "level the playing field" between the two models. The reduction in sample size for the generic dataset was also done to reduce the computational costs. Future work could improve upon this by finding appropriate ways to compare performance for models trained on datasets of different sizes, allowing us to use the full dataset instead of random samples.

We also find this project can be further used to extract other voting behaviors for different levels of demographic groups. For instance, we found mid-aged people would have the most different voter share when surveyed with generic question and specific question. With the combination of two demographics, mid-age, white people would have the most different voter share. We could potentially further analyze this combination, and related findings may be helpful to improve the survey structures for future use.