Predicting Error in US Federal Election Polling

Matthew Burgess

May 2020

Introduction

In Presidential, Senatorial, and Congressional elections across the United States, all campaigns and media outlets rely on polling. Exit polls help determine decisions on election day. Polls of Iowa in presidential primaries help determine the front-runner for a given race. But how often do these polls reflect the actual outcome of elections? More importantly, how wrong are they? My goal is to create a model that predicts the error of a poll from the actual result.

Being able to predict the error of a poll is one of the single greatest indicators of reliability. In being able to predict how far from the actual result of an election a poll is, campaigns and media outlets can more uniformly tailor both political strategy and political coverage.

A model that predicts the error of polls is also useful for pollsters themselves. Polls cost pollsters money, from the overhead of hiring people to conduct the polls, technology, and the number crunching to prepare the data for release. A main driver of this is sample size. Small sample sizes and small error are two important marks to hit for the organizations conducting polls. If we can also find the smallest sample size necessary for an effective poll, we can also provide pollsters with information on running polls efficiently.

In sum, I want to create a model that predicts the error of a poll, and also hopefully explore the data I use to find what the smallest sample size could be that minimizes the error of the poll.

Data Summary and Wrangling

Form of Data:

I plan on using data from the data news site FiveThirtyEight on polls. It is roughly 10000 rows by 25 columns from polls since 1998. Please see the link here to get to the CSV file. It not only includes data of polls of race by party, but also the year, partisan lean of both the pollster and the district, and real percentage of votes obtained by a candidate.

Feature Engineering:

In examining polls and their error, the sample size of the poll itself is most important as an indicator of whether the margin will be significantly large. Therefore, I’ve kept all of the polls regardless of sample size, and try to mitigate any potential effects of incredibly large or small samples.

Here are the descriptions of columns as described directly Fivethirtyeight:

pollno: FiveThirtyEight poll ID number

race: Election polled

year: Year of election (not year of poll)

location: Location (state or Congressional district, or "US" for national polls)

type_simple: Type of election (5 categories, Gov-G, Sen-G, Pres-G, Pres-P, House-G)

type_detail: Detailed type of election (this distinguishes between Republican and Democratic primaries, for example, whereas type_simple does not)

pollster: Pollster name

partisan: Flag for internal/partisan poll. "D" indicates Democratic poll, "R" indicates Republican poll, "I" indicates poll put out by independent candidate's campaign. Note that different sources define these categories differently and our categorization will often reflect the original source's definition. In other words, these definitions may be inconsistent and should be used carefully.

polldate: Median field date of the poll

samplesize: Sample size of the poll. Where missing, this is estimated from the poll's margin of error, or similar polls conducted by the same polling firm. A sample size of 600 is used if no better estimate is available.

cand1_name: Name of Candidate #1. Candidates #1 and #2 are defined as the top two finishers in the election (regardless of whether or not they were the top two candidates in the poll). In races where a Democrat and a Republican were the top two finishers, Candidate #1 is the Democrat and simply listed as "Democrat".

cand1_pct: Candidate #1's share of the vote in the poll.

cand2_name: Name of Candidate #2. Candidates #1 and #2 are defined as the top two finishers in the election (regardless of whether or not they were the top two candidates in the poll). In races where a Democrat and a Republican were the top two finishers, Candidate #2 is the Republican and simply listed as "Republican"

cand2_pct: Candidate #2's share of the vote in the poll.

cand3_pct: Share of the vote for the top candidate listed in the poll, other than Candidate #1 and Candidate #2.

margin_poll: Projected margin of victory (defeat) for Candidate #1. This is calculated as cand1_pct - cand2_pct. In races between a Democrat and a Republican, positive values indicate a Democratic lead; negative values a Repubican lead.

electiondate: Date of election

cand1_actual: Actual share of vote for Candidate #1

cand2_actual: Actual share of vote for Candidate #2

margin_actual: Actual margin in the election. This is calculated as cand1_actual - cand2_actual. In races between a Democrat and a Republican, positive values indicate a Democratic win; negative values a Republican win.

error Absolute value of the difference between the actual and polled result. This is calculated as abs(margin_poll - margin_actual)

bias: Statistical bias of the poll. This is calculated only for races in which the top two finishers were a Democrat and a Republican. It is calculated as margin_poll - margin_actual. Positive values indicate a Democratic bias (the Democrat did better in the poll than the election). Negative values indicate a Republican bias.

rightcall: Flag to indicate whether the pollster called the outcome correctly, i.e. whether the candidate they had listed in 1st place won the election. A 1 indicates a correct call and a 0 an incorrect call; 0.5 indicates that the pollster had two or more candidates tied for the lead and one of the tied candidates won. comment Additional information, such as alternate names for the poll.

The data is organized such that the column “cand1_pct” is always a Democratic race for general presidential election polls and “cand2_pct” is always a Republican race for general presidential election polls. When the error is calculated as a difference between the two rows actual results respectively, a positive value for the column “margin_actual” indicates a Democratic victory, and a negative value for the same column indicates a Republican victory. I have also deleted the “comment” column as well as removed all partisan pollsters as indicated by the “partisan” column, of which there were few.

Feature engineering to prepare for regression included the creation of dummy variables for all categorical columns, and the deletion of duplicate columns.

Missing and NaN Data

Data exploration and wrangling revealed that the only missing data was for races where there was a third candidate in the race. These cells contained NaN or no value. To complete the regression problem, I filled the NaN and missing values with zeroes. This reflects that there is either no third candidate or the percentage value of any other candidates in the race approach zero anyway. This makes sense in a real-world application because the United State's two-party system only allows for third-party to obtain a de minimus percentage of the vote. This has notable exceptions in key presidential races throughout history (Ross Perot), but those are not reflected in this data. As such, I will delete duplicate columns and fill NaNs with 0.

Exploratory Data Analysis and Inferential Statistics

Exploratory Data Analysis

I am most interested in exploring the error associated with polls as a means of determining reliability. I will explore the error, including histograms and scatter plots of various factors and the error. Because the error is a calculation of margins of error, I also will explore the margins of actual and predicted outcomes to see if there are any outlying error values.

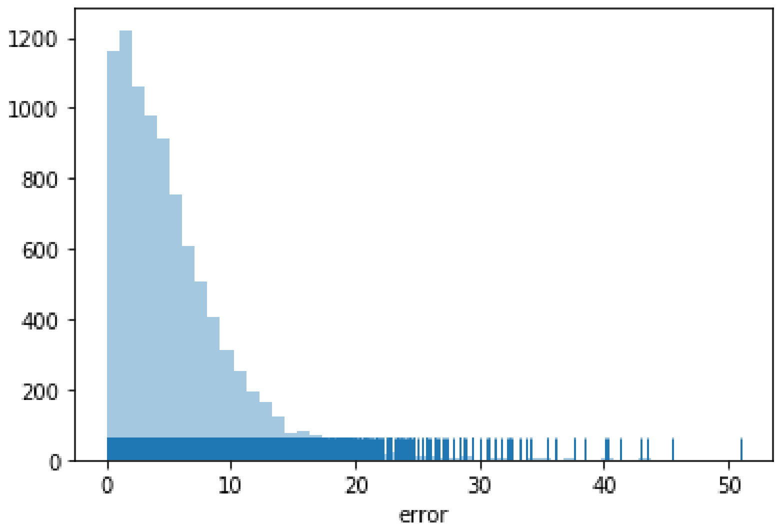

Figure 1-1: Distribution of error with Kernel Density Estimate

According to the distribution of error in Figure 1-1, the error is mostly contained to values below 10 percent, with a majority of the values being under 10 percent. The Kernel Density Estimate also shows that there are a handful of values which have an error of greater than 25 percent. These large sample sizes may be outliers or have excess weight on the error, but exploring further should help to determine if there is collinearity.

The data encompasses polls over a span of 20 years, from 1998 to 2018. Because the techniques and methods of polling have changed significantly (direct voter contact vs. phone calls, internet access, hyperpolarization), there could be large differences in sample size based on the year of the poll.

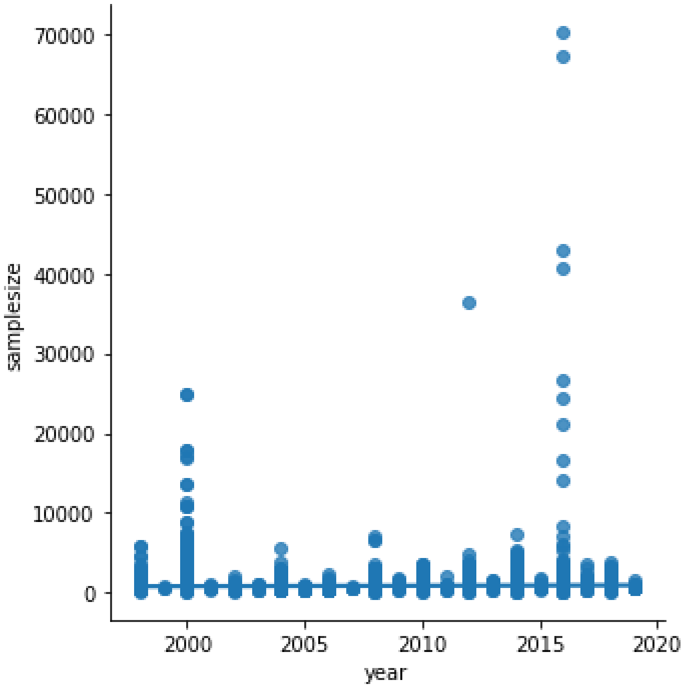

Figure 1-2: Sample Size Distribution Over Time

Figure 1-2 shows that more recent polls have had larger sample sizes, particularly in 2016. It also shows that there are spikes in the number of polls every four years, which is unsurprising since that coincides with presidential election years. There is a significant amount of polling for 2016, and they are quite large polls. This makes sense given the enormity of the election.

Bias is a factor that comes up in many polls. Whether the pollster is partisan or the number of people contacted (sample size) is large or small, there are a variety of factors that go into assessing the bias of a poll. In this data set, the bias is calculated as:

Margin of Actual Election - Margin of Poll Prediction

Because the metrics in the above equation are also used in the calculation of error, it is important to assess if the bias is skewed in any particular direction, which would also indicate a partisan skew.

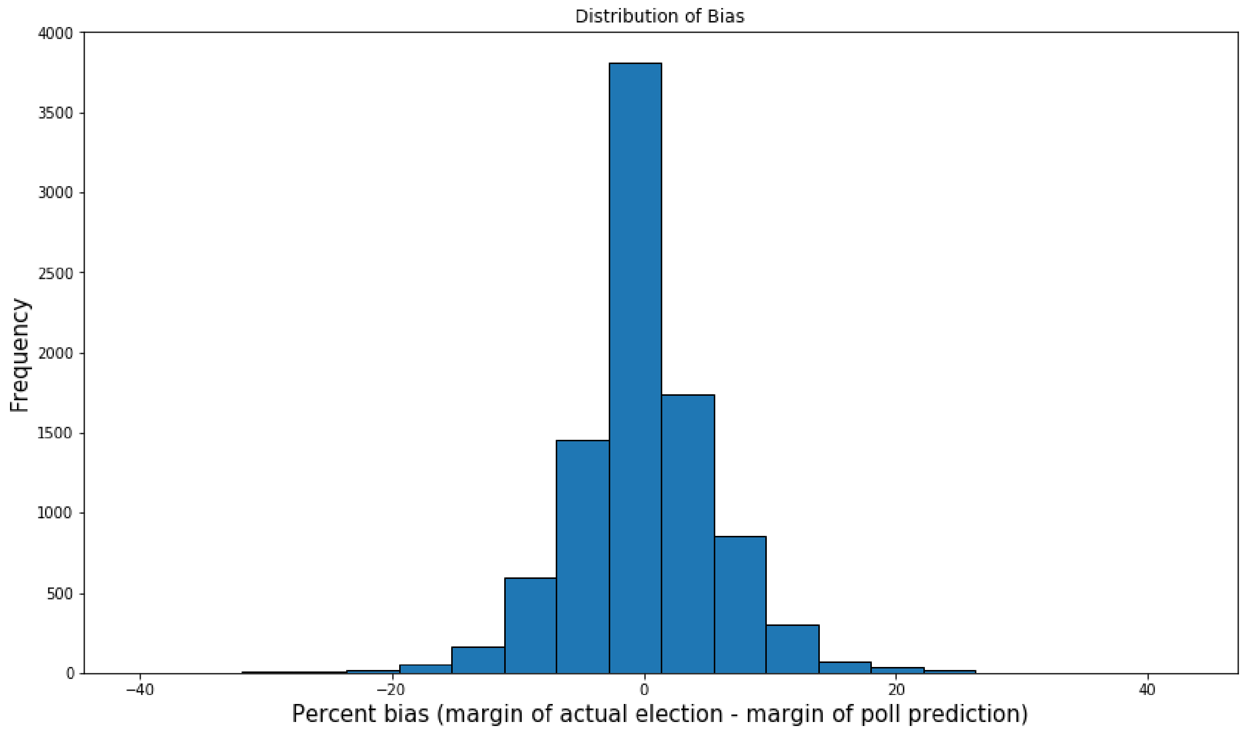

Figure 1-3: Distribution of Bias

As Figure 1-3 shows, the distribution of the bias of the polls appears to be unimodal and symmetric, approximately normal.

Collinearity with any particular feature may result in overfitting. The most concern with this is in the sample size feature, where extremely large or small values may result in outsize error. As figures 1-4 shows below, there appears to be no collinearity between the sample size and error. Similarly, figure 1-5 shows the error plotted as the scatter plot of the functions that make it up. If there were a line running through the graph in 1-5, the error is the distance to that line for each point. There are few points significantly orthogonally far away from that line.

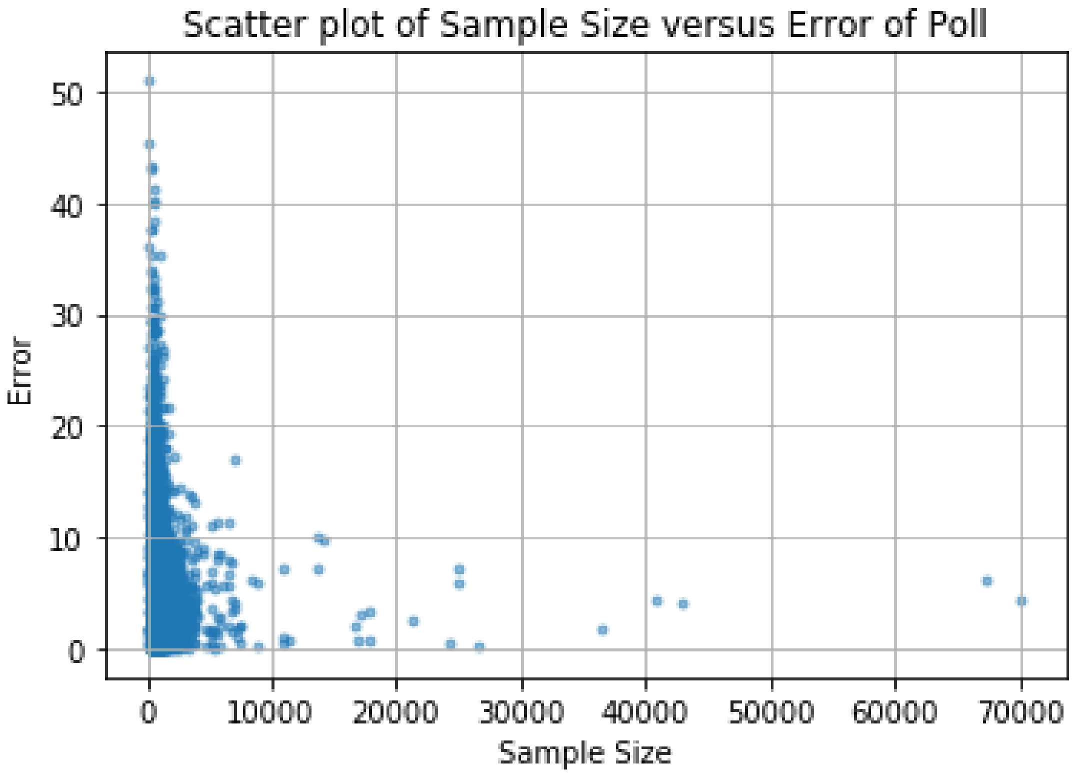

Figure 1-4: Sample Size vs. Error

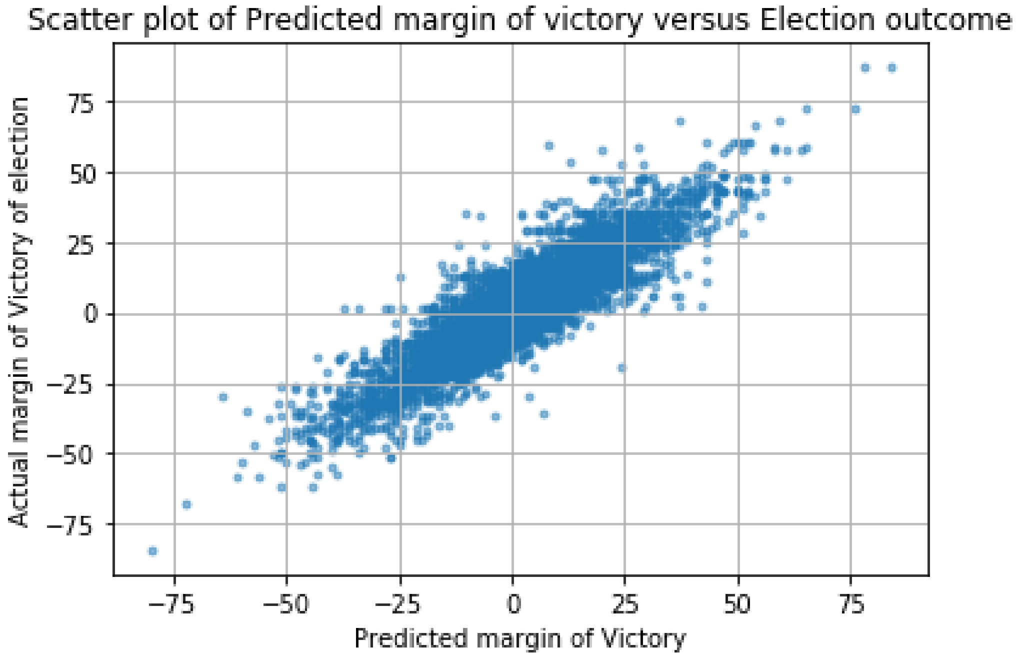

Figure 1-:5 Scatter Plot of Predicted margin of victory versus Election outcome

Machine Learning and Model Selection

I am interested in creating a regression that predicts the error of a model. Because I suspect some features may have a greater impact on the result and I want to use the most up-to-date methods, I will start by using the following methods, using a testing data of 33%.

1) Basic Linear Regression

Doing a basic linear regression on structured data is the best place to start. While I don't expect this to be the best starting point, linear regression will help reveal if we should be doing lasso or ridge regression due to collinearity with particular features, particularly the sample size of the poll. I believe that this could be overfit.

2) XGBoost Regression

XGBoost is a Decision Tree algorithm that uses gradient boosting for decision tree algorithms to minimize loss among new models. In researching machine learning algorithms to use, XGBoost has been used as an industry standard for being computationally fast and producing an effective model. It also executes and adds new models of the decision tree forest under the hood, resulting in more accurate results.

3) Lasso Regression

Lasso Regression is an effective algorithm for when we believe a particular feature may have an outsized weight on the regression. Lasso does this using an absolute value function within its regularization parameter, causing a more intense drop for outsized weights. In this way, it 'punishes' high values. If sample size has increased collinearity with the outcome, then Lasso should be able to handle this and predict the error.

4) Ridge Regression

Ridge,like Lasso regression, also allows us to mitigate outsized variables and overfitting. It penalizes high values by using a squared magnitude rather than an absolute value. Using ridge in comparison with a Lasso will provide insight into which method is more effective at reducing model complexity.

Results

I trained the above models on our training data, and then had it predict the error using the testing data. The following scores were found for the models' predictions compared to the actual values.

Method	**R-Squared Score**	**Mean Squared Error**	**Mean Absolute Error**
Linear Regression	**0.17545737688184448**	**18.73358721911903**	**3.1206837093578743**
XGBoost	**0.8274592710021751**	**3.920120931174652**	**0.9238756385858912**
Lasso	**0.07534753228877467**	**21.00808031698279**	**3.3709132867746283**
Ridge	**0.21706739446888657**	**17.78820868828203**	**3.0539818617526806**

Figure 2-1: Table of Results from Test Data

In figuring out which algorithm best predicts the data, it is clear that XGBoost Regressor provides the best prediction of the error in election polling. Tt has a higher R-Squared value than lasso or ridge, and it also provides the lowest mean absolute error of 0.92. As a value of prediction of error, this means that the model is on average 0.92. Also, the 'error' value that we are predicting is a percentage, so our MAE of 0.92 means that we are usually within 1% for any value. XGBoost clearly is the best predictor.

This also reveals that there does not seem to be an overfitting issue, as our measures of R-square, Mean Squared Error, and Mean Absolute Error of the training data are close to those of the testing data. This also supports that Lasso and Ridge are not the most effective algorithms, as we do not have an overfitting or collinearity issue.

In applying this to a real-life political situation, reacting to new polling data and knowing how far we expect it to be off from the actual election result can have enormous political consequences. While this is just a start and there are various socio-political factors to also take into account, using this method of regression to predict error can be a great starting point, and provides many questions about future inquiry. We were ultimately able to predict the error of a poll from its general election outcome accurately 82% of the time, and our average error on that estimate was less than one percentage point.