

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 $\text{cand1_pct} - \text{cand2_pct}$. In races between a Democrat and a Republican, positive values indicate a Democratic lead; negative values a Republican 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 $\text{cand1_actual} - \text{cand2_actual}$. 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 $\text{abs}(\text{margin_poll} - \text{margin_actual})$

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 $\text{margin_poll} - \text{margin_actual}$. 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.

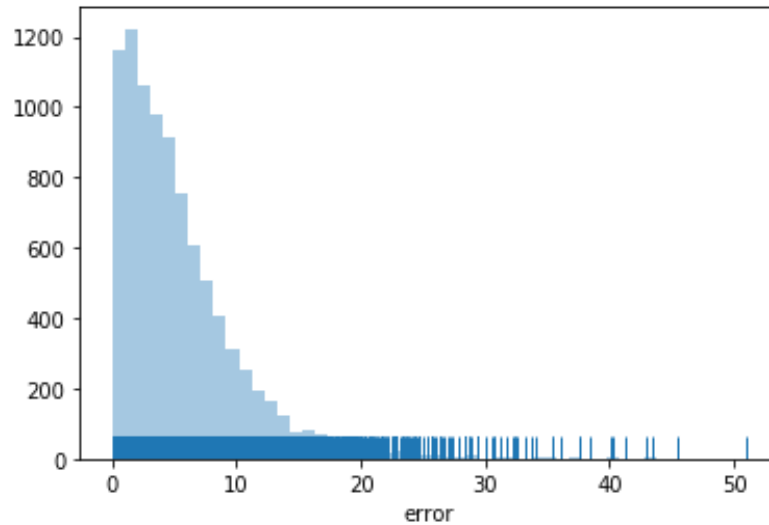


Figure 1-1: Distribution of error with Kernel Density Estimate

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.

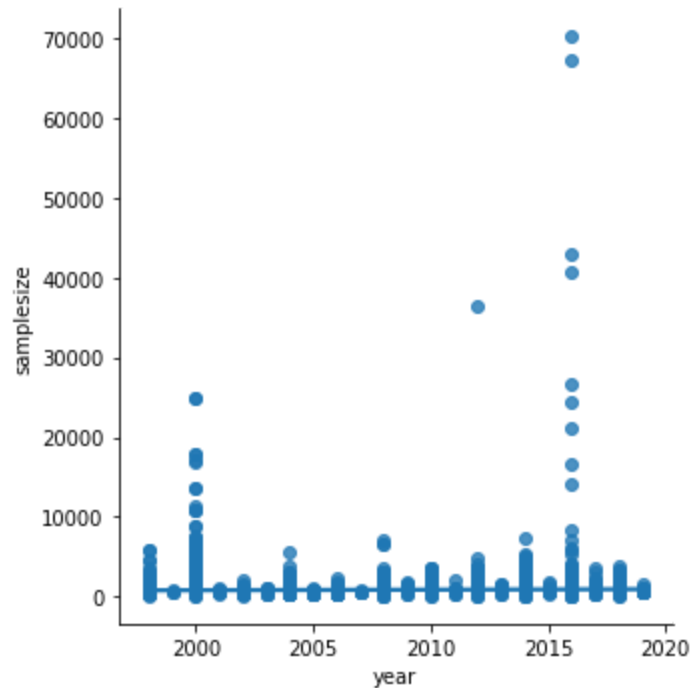


Figure 1-2: Sample Size Distribution Over Time

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:

$$\text{Margin of Actual Election} - \text{Margin of Poll Prediction}$$

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.

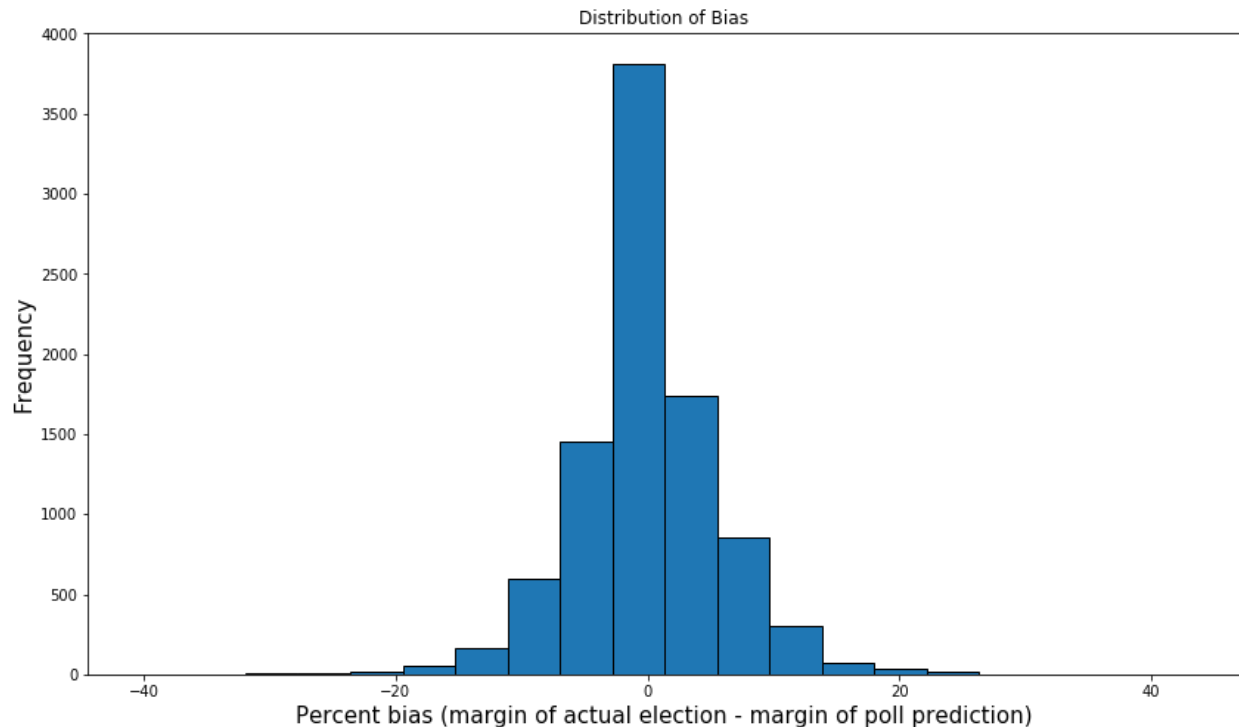


Figure 1-3: Distribution of Bias

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.

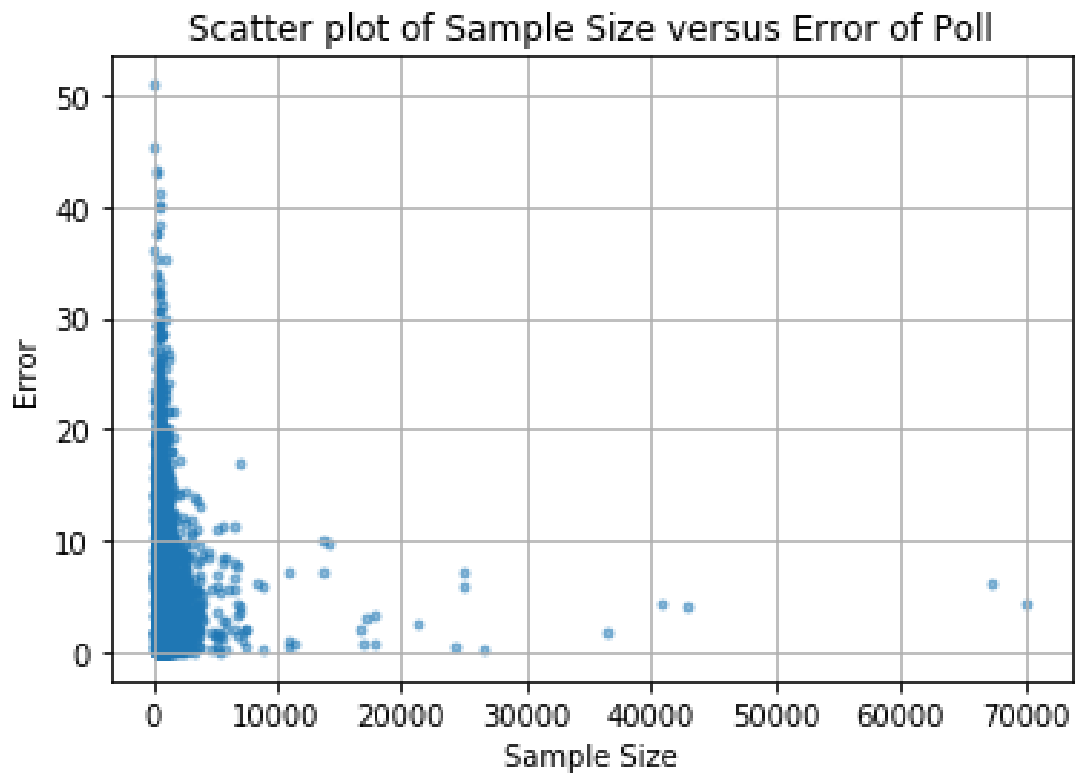


Figure 1-4: Sample Size vs. Error

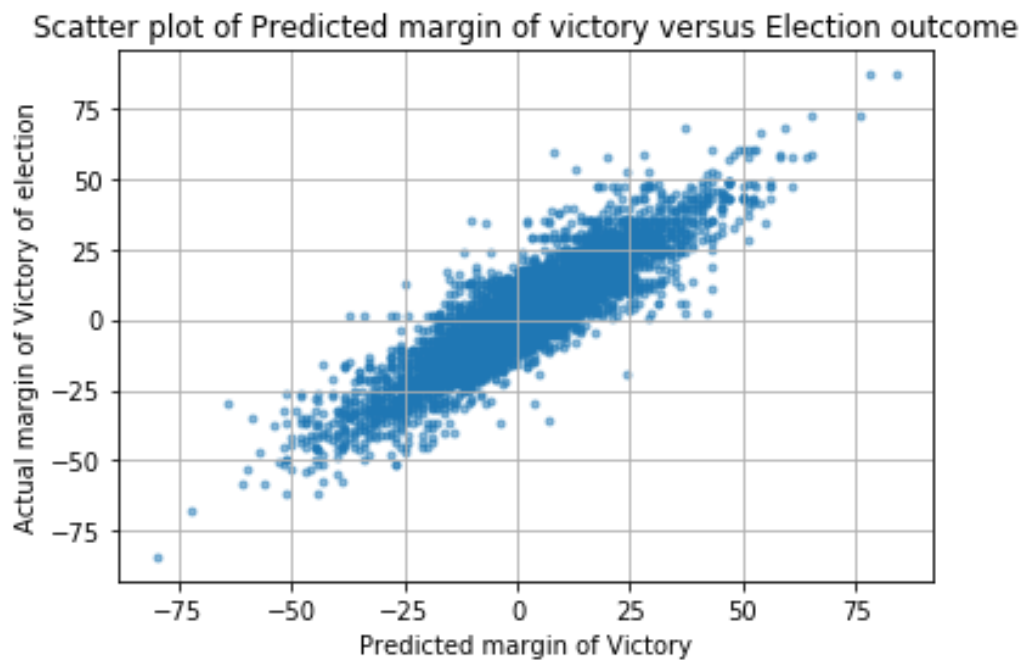


Figure 1-5 Scatter Plot of Predicted margin of victory versus Election outcome

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 $\text{cand1_pct} - \text{cand2_pct}$. In races between a Democrat and a Republican, positive values indicate a Democratic lead; negative values a Republican 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 $\text{cand1_actual} - \text{cand2_actual}$. 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 $\text{abs}(\text{margin_poll} - \text{margin_actual})$

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 $\text{margin_poll} - \text{margin_actual}$. 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.

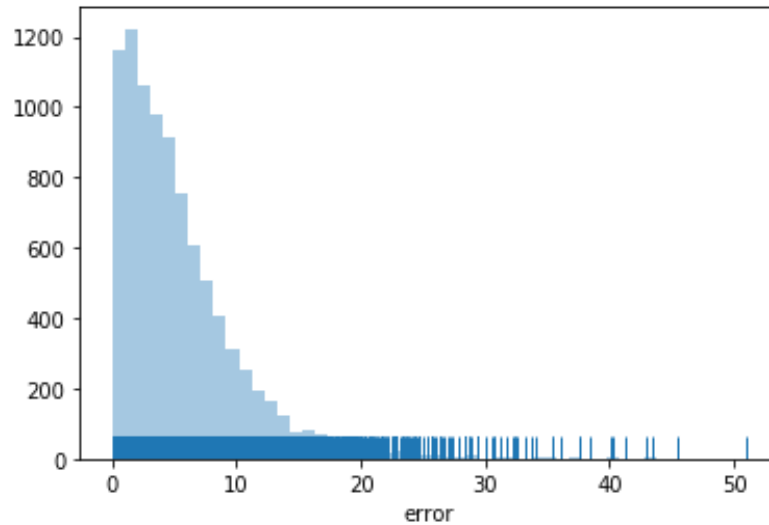


Figure 1-1: Distribution of error with Kernel Density Estimate

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.

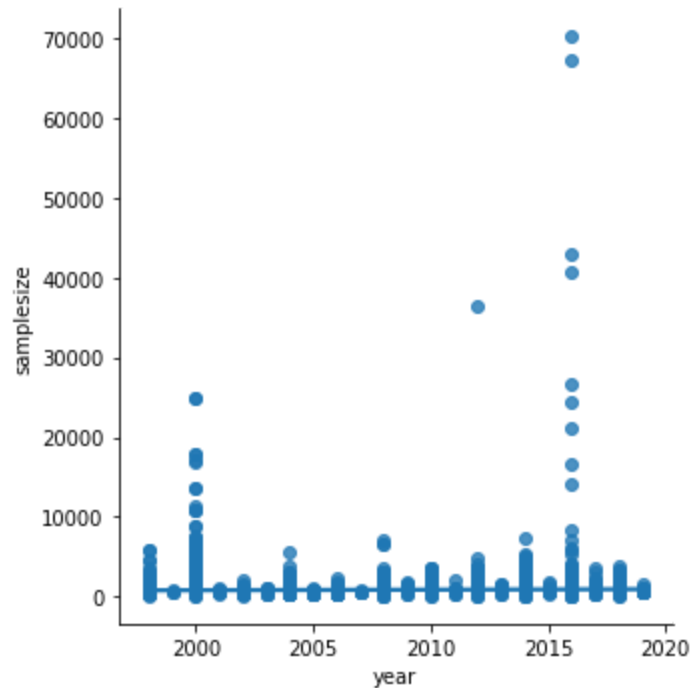


Figure 1-2: Sample Size Distribution Over Time

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:

$$\text{Margin of Actual Election} - \text{Margin of Poll Prediction}$$

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.

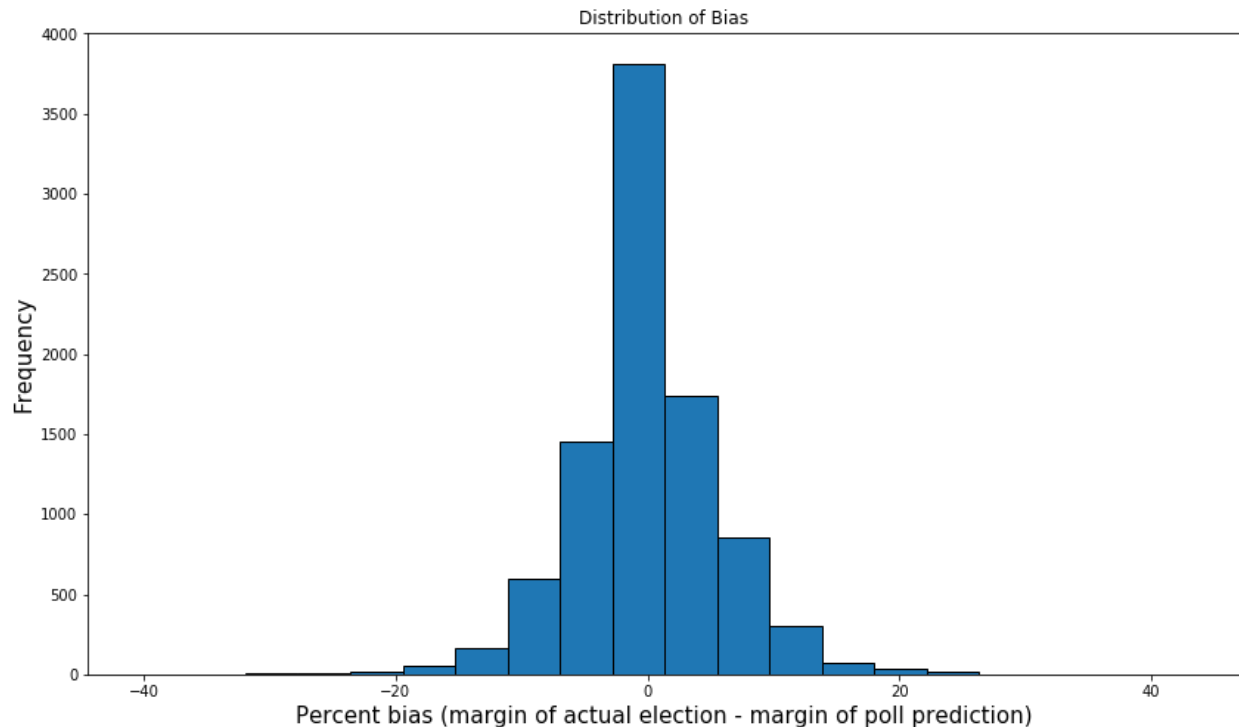


Figure 1-3: Distribution of Bias

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.

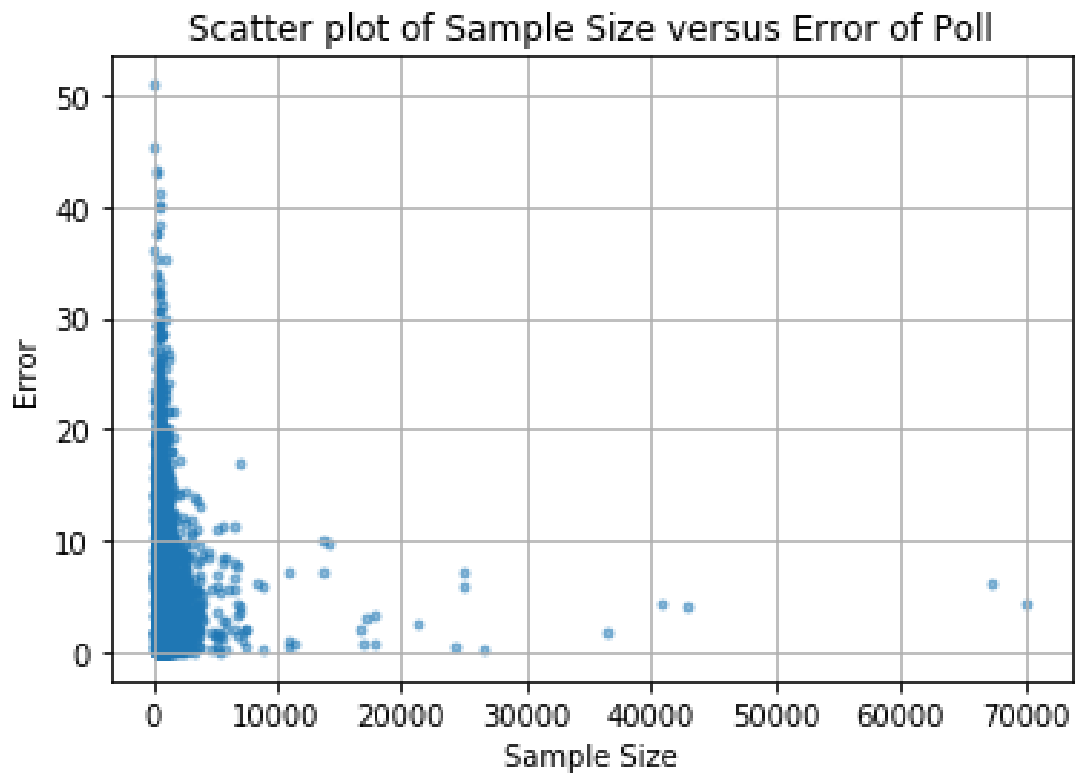


Figure 1-4: Sample Size vs. Error

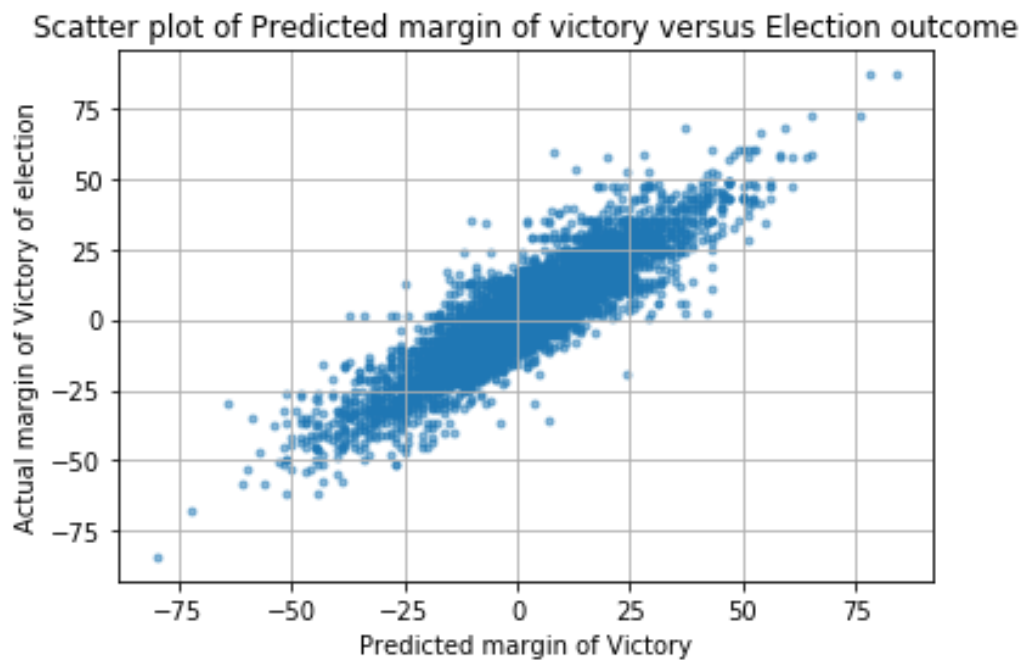


Figure 1-5 Scatter Plot of Predicted margin of victory versus Election outcome