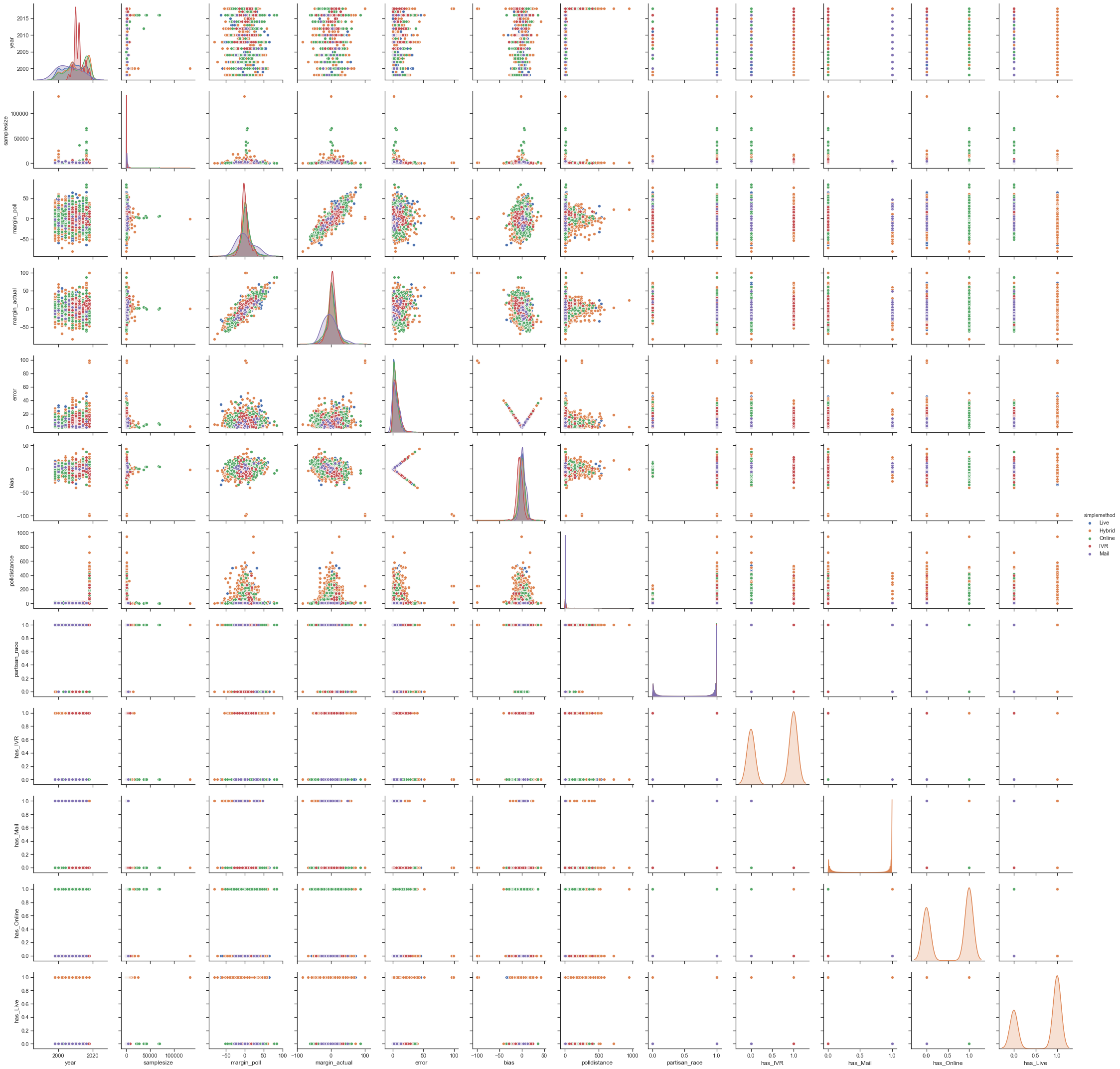
**Capstone Project I: Statistical Data Analysis**



Seaborn pairplot of year, sample size, predicted/actual margin, error, bias, and poll distance, colored by poll type.

The target dependent variables in the FiveThirtyEight dataset are bias and error. Both variables negatively corelate to sample size.

A picture containing text

Description automatically generatedA picture containing text

Description automatically generated

Matplotlib plots of bias and error against poll sample size.

Determining the relationship between error/bias and polling sample size was complicated by the fact that sampling size described a maximum error, but did not restrict the minimum. In other words, it was possible to sample 50 people and make an error-free poll, but a poll with 7500 respondents was very unlikely to be badly wrong. To account for this, I binned the data into 50-person groups and took the maximum of each.

This produced the following plots:

A close up of a map

Description automatically generatedA close up of a map

Description automatically generated

We can see that maximum bias and error have a roughly linear relationship with sample size up to a point. Increasing sample size beyond this threshold does not significantly improve accuracy.

There’s clearly significant variation in polling, but we can estimate a linear regression for these datasets regardless.

A close up of a map

Description automatically generatedA screenshot of a cell phone

Description automatically generated

A simple regplot of sample size against max error up to 4000 people has a p-value of 0: sample size is a statistically significant predictor of max error. R-squared of 0.66, F-stat of 141.

Plotting values against residuals indicates a linear model may not be the best fit for this data.

A close up of a map

Description automatically generatedA close up of text on a white background

Description automatically generated

Adjusting with np.log(error) produced a worse model, but with a better fitted values/residuals plot. R-squared of 0.60, F-stat of 114.

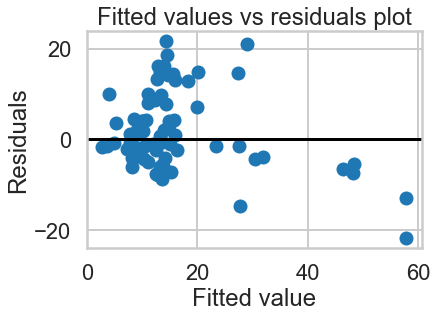
Standard deviation of error of previous polls by a pollster is a strong indicator of their future results, as seen here:

A close up of a piece of paper

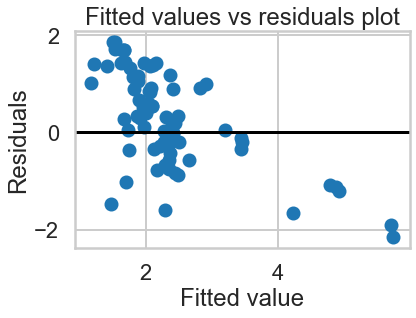
Description automatically generated

Plugging in mean standard deviation of the pollster for the OLS model brought us to R-squared of 0.832, F-stat of 184. I can add mean error of pollster to raise R by 1%, so I think that mostly just overfits.

This model predicts the majority of maximum polling error based on previous performance by a pollster and sample size of poll.



Testing the log of error against the samplesize+std regression gives a R-squared of 0.859, F-stat of 226. In the future, I’ll examine expanding this out to predict all error, not just maximum error. This is the best model I’ve found yet for predicting error in polling.



I’ve experimented with other methods, such as state and poll type. So far nothing’s worked as well.

A screenshot of a cell phone

Description automatically generatedA screenshot of a cell phone

Description automatically generated

As seen here, polling type is a strong indicator of bias, but it fares poorly when predicting error, because positive/negative biases in hybrid, live, and online polls stop canceling each other out.

One of the persistent questions I’ve had when working on this dataset is: Why does the time before the election appear to negatively correlate with the error of the poll? In other words, why are polls released close to the election less accurate than polls released far from the election? This is a graph of poll distance versus error, with darker points representing smaller sample sizes.

A close up of a map

Description automatically generated

I believe that a large part of the reason that we see a lot of high-error polls close to elections is that there are a lot of fast, low-sample-size polls rushed out immediately prior to an election. There are still a lot of fast polls that get things right, either through luck or through herding, but on average they do worse.