Election Forecasting Using Linear Regression

by Gene Lam

Introduction

Regression analysis is a method of identifying relationships between two or more variables. It allows you to determine which factors can impact a dependent variable and by how much. This makes regression analysis incredibly useful for creating models that depict relationships between multiple agents. A correlation coefficient of -1 or 1 indicates a strong association between the observed variables. However, note that just because there may be a relationship, that does not necessarily indicate causality. For example, the divorce rate in Maine seems to decrease at the same rate as per capita consumption of margarine, but that does not mean consuming margarine affects marital satisfaction. This distinction is humorous, but has important implications in scientific literacy as people use these relationships to justify a link between vaccines and autism. In reality, there are multiple environmental factors at play, including a broadening of diagnosing people on the autism spectrum or changes in generational lifestyles.

The same precaution must be applied to election modeling and election forecasting. These models are used to inform the public on what to expect, and provide meaningful insight into effective strategies for politicians and their campaigns. However, the models are not perfect as they rely on limited data from the past. In 2016, Hillary Clinton was predicted to be the first female President in a majority of notable forecasts. However, the 2016 election was unique in that the Republican candidate, unlike in previous elections, ran on the platform of being an outsider. Donald Trump ran as an unapologetic businessman ready to disrupt the political machine in Washington. The results from the 2016 election proves just how unreliable models can get when you introduce an outlier to a sea of data.

Methodology

To explore election modeling, we attempt to create an election model using regression to predict who will win the 2020 election. In his blog post on modeling the Poland Presidential election, data scientist Cezary Klimczuk, argued two key parameters for predicting election results: voter-base engagement and third party engagement. Klimczuk argues that in a re-election campaign, Presidential candidates focus on two goals: retain their voter base and persuade third party voters to join their base. We will adapt that to the regression model in the US Presidential elections:

$$\Delta Y = \beta_0 x_0 + \beta_1 x_1 + \beta_2$$

ΔY = change in % total votes

β_0 = voter-base correlation coefficient

β_1 = third party correlation coefficient

β_2 = constant

The method of least-squares was used to calculate the correlation coefficients and intercept constant. Recall the formula for determining these values:

$$a_0 = \frac{\left(\sum_{i=1}^{m} x_i^2\right)\left(\sum_{i=1}^{m} y_i\right) - \left(\sum_{i=1}^{m} x_i y_i\right)\left(\sum_{i=1}^{m} x_i\right)}{m\left(\sum_{i=1}^{m} x_i^2\right) - \left(\sum_{i=1}^{m} x_i\right)^2}$$

$$a_1 = \frac{m\left(\sum_{i=1}^{m} x_i y_i\right) - \left(\sum_{i=1}^{m} x_i\right)\left(\sum_{i=1}^{m} y_i\right)}{m\left(\sum_{i=1}^{m} x_i^2\right) - \left(\sum_{i=1}^{m} x_i\right)^2}.$$

To determine the voter-base correlation coefficient, a candidate's first election turnout was plotted against their re-election turnout. A dataset from the MIT Election Lab gives us the percentage of votes each Presidential candidate received in 50 states from 1976 to 2016 (see **Appendix: Figure**

1). It seems like there is a strong positive correlation of 0.7 between voting preferences in a President's first race and re-election race, which makes sense as people tend to support the candidate that they voted for in previous elections. The red dots represent Republican candidates and the blue dots represent Democratic candidates, and the voter trends are the same for both Democrats and Republicans so a general trendline can be plotted. Since the dependent variable is ΔY, the percent change in total votes between elections, a President's percentage of votes in their election is subtracted from their percentage of votes in their re-election. The scatterplot below depicts the relationship between a President's first election percent of votes and the percent change in their re-election.

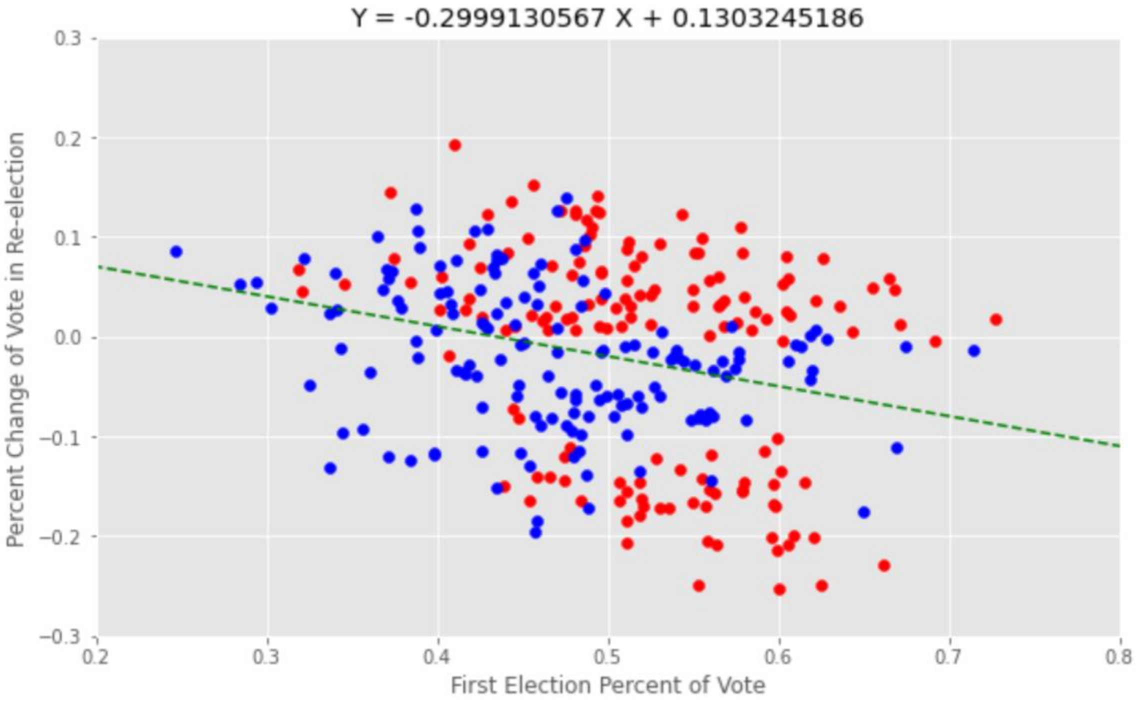

Figure 1: First Election Percent of Vote vs. Percent Change of Vote in Re-Election

This tells us that a Presidential candidate who gains a higher percent of vote in their first election receives slightly less votes in their re-election. The reason for this phenomenon is unknown, but perhaps a President with a high percentage of votes is given high expectation from society, which backfires in the re-election. It could also be that Presidents who win high by high margins tend to overpromise and underdeliver. The line of best fit for the voter-base parameter is:

$$\Delta Y = -0.2999130567x_0 + 0.1303245186$$

where x_0 is the percentage of votes in a President's first election

The third party contribution is then calculated by finding the relationship between third party turnout in a candidate's first election and the percent change in votes in their re-election.

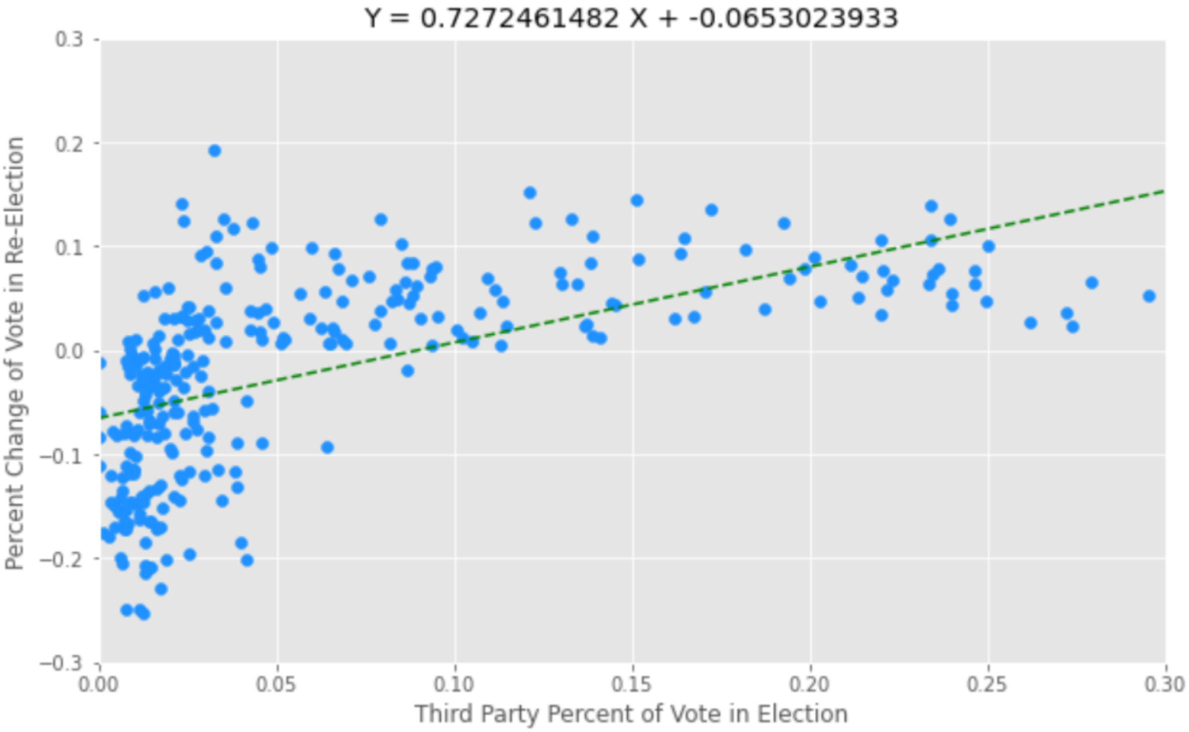

Figure 2: Third Party Percent of Vote in Election vs. Percent Change in Vote in Re-Election

Notice that the line of best fit does not properly match the data linearly and appears logarithmic in nature. Due to this, a greater skepticism is applied to the third party contribution despite the correlation coefficient indicating a strong relationship. The data is skewed so that a low third party turnout in a candidate's election results in losing voters in the re-election whereas a candidate with a higher third party percentage of voters gains votes. Intuitively, a President who in their election has a high percent of third party voters will have more opportunities to sway these voters in their re-election. The less third party voters there are, the more people have already made their mind up about whether to elect someone in their re-election. The third party contribution line of best fit is:

$$\Delta Y = 0.7272461482x_1 - 0.0653023933$$

where x_0 is the average voter base turnout in a candidate's first election

The two linear regression equations were added together to come up with the final model:

$$\Delta Y = -0.2999130567x_0 + 0.7272461482x_1 + 0.0650221253$$

The percent of votes for Donald Trump in his first election was found as well as the percent of third party votes in the same election to determine the change in percentage of votes Donald Trump can expect in the 2020 election.

Results

The election forecast showed that Donald Trump lost at 45.97% and Joe Biden won the 2020 election with a predicted vote of 49.81%. The predicted vote for Joe Biden was calculated by performing a linear regression to predict the number of voters voting third party in the re-election (see **Appendix: Figure 2**). This predicted value was subtracted from 100, along with Donald Trump's predicted value. Since the election past, the absolute and relative errors were calculated accordingly:

Table 1: 2020 Presidential Election Predicted and Actual Results

	Vote %	Predicted Vote %	Absolute Error	Relative Error
Donald J. Trump	46.8%	45.99%	1.113	0.024

Joe Biden	51.3%	49.81%	1.292	0.025

The predicted win/lose margin, -3.82%, is also in line with other Presidential elections:

Table 2: Presidential Election Actual Margin of Win

Incumbent	Challenger	Margin of Win
George H.W. Bush	Bill Clinton	-5.6%
Bill Clinton	Robert Dole	8.5%
George W. Bush	John Kerry	2.4%
Barack Obama	Mitt Romney	3.9%
Donald J. Trump	Joe Biden	-4.5%

Limitations

Elections are complex systems that have multiple factors to take into consideration. As the primary candidates optimize their campaign strategies to maintain their respective bases and win over third party voters, these races can get incredibly tight, making it difficult to create these models. The one we just created is overly simplified, and a rough estimation. It does not take into account public opinion, turnout, trust in institutions, campaigning and marketing, etc - all of which have been impacted by recent events including the Black Lives Matter protests, the pandemic, and the actions of Donald Trump. Elections are determined by the people and we all have our unique experiences, values, and beliefs that go into who we vote for on the ballot.

There are also limitations in using linear regression such as incomplete or lack of data. From the third party example, it was observed that the graph was not linear and therefore, another type of analysis could be used.

Application in the Georgia Senate Regular Election Runoff

Due to Georgia's Election laws, the Georgia Senate races will be held in a runoff on January 5. The same methodology was applied to the Jon Ossoff v. David Purdue Senate Race Runoff. A regression model could not be performed on the Kelly Loeffler v. Reverend Raphael Warnock Senate Special Runoff as that race included multiple Democrats, Republicans, and third party candidates. It would be difficult to discern where the votes for those candidates would go if not to the respective candidates for each party.

Although the same methodology was used, a different dataset was used to create the models, one specifically for US Senators from 1976-2018. After the voter-base and third-party correlation coefficient was calculated, the two equations were added up, average voter-base and third-party turnout was plugged in, and the election results were calculated. Here are the results:

Table 3: Predicted Senate Runoff Regular Election

David Purdue	48.894%
Jon Ossoff	50.054%

This was surprising as David Purdue received more votes than Jon Ossoff in the November 2020 race. It could be that this methodology skews against incumbents, and that would be something worth examining in the Presidential use case as well. However, the margins are so tight (within +2%) that it could go either way. This is in line with news coverage on the state of Senate race in Georgia. Note that the Georgia regular election model is more prone to error because the dataset was generalized to votes in Georgia, as a whole, rather than each county in Georgia. Additionally, Senate runoff cases are rare so races where an incumbent was defending their position from a challenger was included. As was stated before, outlier cases - cases where the situation is unusual (like during a pandemic, when political tensions are high, when people do not trust elections) - make it difficult to apply reliable and robust models.

Conclusion

In this paper, linear regression was performed to forecast the results of the November 2020 election. The line of best fit was calculated using the least-squares method, and used to estimate voter turnout and third party turnout. Limitations in linear regression were discussed, as well factors that could sway voters. After the experiment was successful in determining the results of the November 2020 election, a model was created to predict the results of the upcoming Senate runoffs.

Works Cited

Klimczuk, C. (2020, 07 15). *How I predicted the election result with simple Linear Regression*. towards data science. Retrieved 11 10, 2020, from https://towardsdatascience.com/how-i-predicted-the-election-result-with-simple-linear-regression-e54c6c196239

MIT Election Data and Science Lab (Massachusetts Institute of Technology). (2019, 04 24). *U.S. President 1976–2016*. U.S. President 1976–2016. Retrieved 20 11, 2020, from doi:10.7910/DVN/42MVDX

MIT Election Data and Science Lab (Massachusetts Institute of Technology). (2019, 04 24). *U.S. Senate 1976–2018*. MIT Election Data and Science Lab. Retrieved 11 12, 2020, from doi:10.7910/DVN/PEJ5QU

The New York Times. (n.d.). *Presidential Election Results: Biden Wins*. The New York Times. Retrieved 12 01, 2020, from https://www.nytimes.com/interactive/2020/11/03/us/elections/results-president.html

Appendix

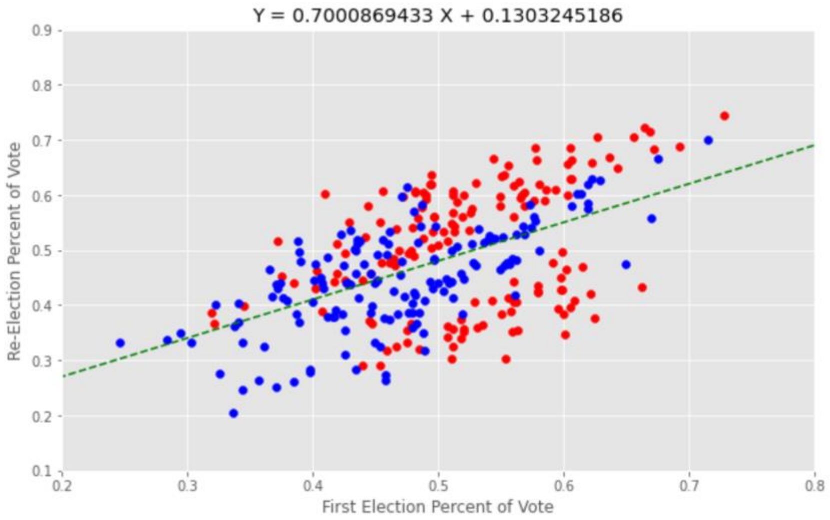

Figure 1: Presidential Election - First Election Percent of Vote vs. Re-Election Percent of Vote

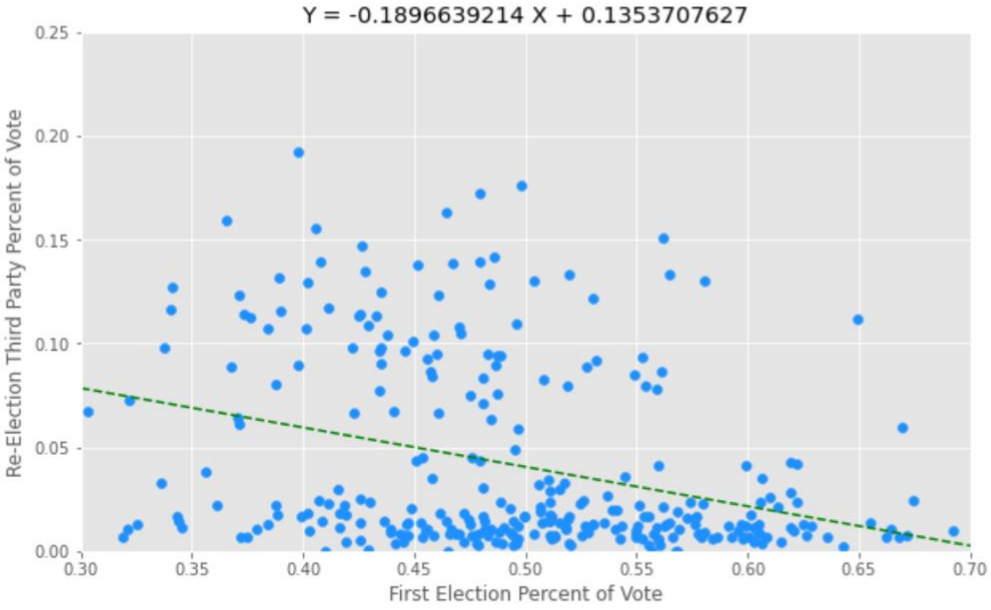

Figure 2: Presidential Election - First Election Percent of Vote vs. Re-Election Third Party Percent of Vote

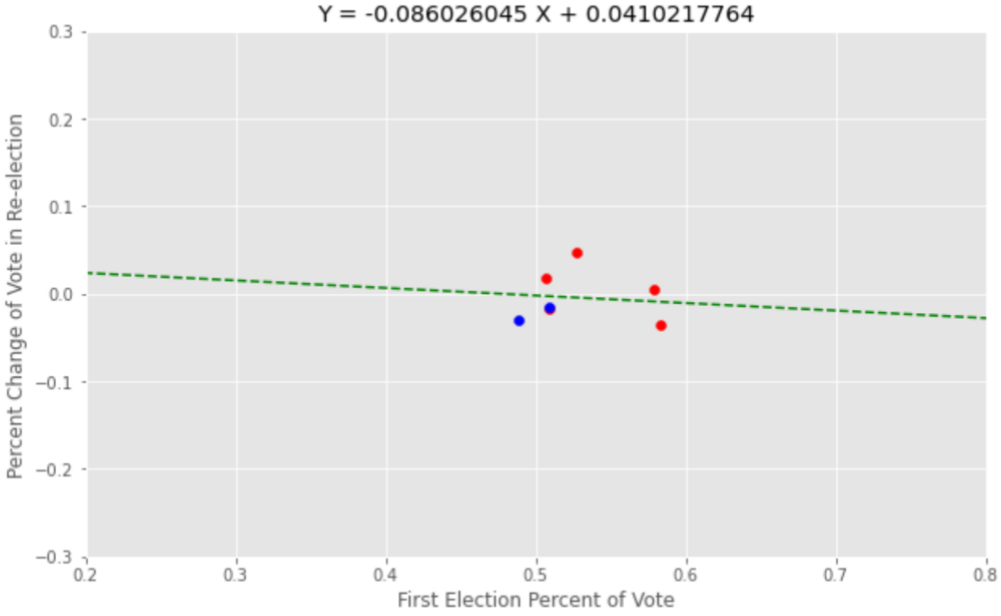

Figure 3: Senate Runoff Election - First Election Percent of Vote vs. Percent Change of Vote in Re-Election

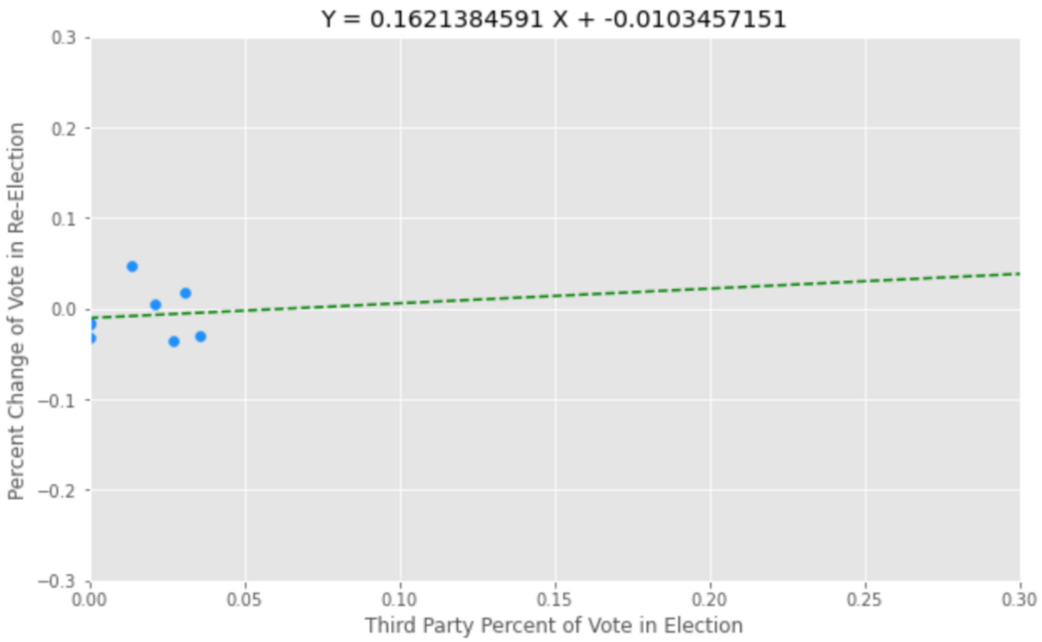

Figure 4: Senate Runoff Election - Third Party Percent of Vote in Election vs. Percent Change of Vote in Re-Election

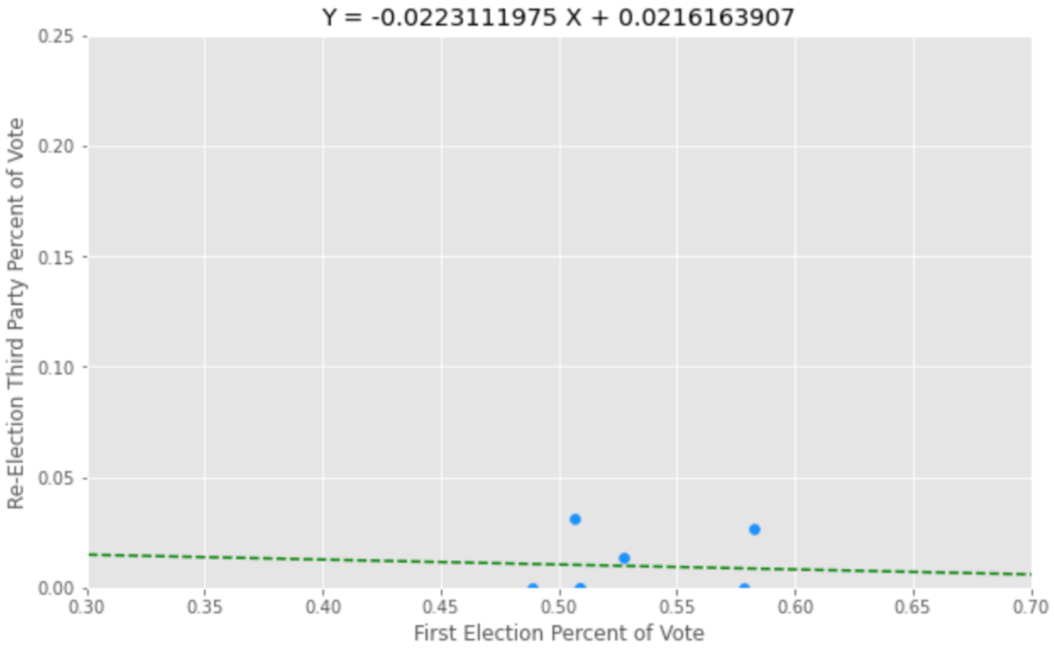

Figure 5: Senate Runoff Election - First Election Percent of Vote vs. Re-Election Third Party Percent of Vote