**Modelling Federal Campaign Fundraising Trends**

**1: Introduction**

I took AP US History last year, and I’m taking AP US Government and Politics this year. As an avid fan of history and social sciences, I had always been intrigued by all the intricacies of the US government system. Unfortunately, the US government was not always portrayed in the most positive light throughout must of history; it has often been deemed as “corrupt”. Sometimes, the corruption is evident – for example, there were a series of major scandals during the Grant administration, and other incidents such as the Watergate Scandal visibly weakened public trust. In contemporary times, outright scandals have been less common, but one of the biggest charges that citizens have on government corruption is the campaign financing status quo.

The US government is elected by the people, so politicians have to campaign to obtain or maintain positions in office. Because elections are hugely publicity-based, more often than not, whichever candidate spends more money on their campaign happens to be the victor. Recognizing this, big corporations and other special interests often donate to support political campaigns, in exchange for promotion of their concerns in the case that the candidate wins. In recent years, critics have held more firmly than ever that the current campaign finance system is corrupt. After Trump took office, the public has questioned even the integrity of our system even more. Hence, in the midst of this turmoil, I wish to evaluate the statistics in federal campaign financing and interpret what these trends may reveal about our federal government.

Math and government did not always go hand in hand for me. At first, they seemed like irreconcilably different disciplines – nothing seemed to be shared between them. However, in government class, we recently learned about the efficiency gap formula, a math equation measuring the extent of gerrymandering. This manifestation of math in politics fascinated me, and let me realize that math could be used to analyze political trends. Hence, in this investigation I will look at the extent and way in which outside money is funneled into US federal politics through the campaign process.

**2: Aim and Approach**

In this paper, I will be modeling the trend of campaign financing in Senate elections since 2008. The year 2008 was chosen because it is the earliest year on which I can find data regarding campaign finance costs. I will be tracking Senate races because it is where most federal campaign money goes – Senate races are far more expensive than races for the House of Representatives, and they also happen as frequently as every 2 years, which is twice as frequent as presidential elections. As is often seen in the news, the amount of money poured into campaign financing has reached all-time highs in 2020, and is projected to continue increasing in the future. I will be trying to find a model that can best describe this increase, and possibly project how much more money will be spent on campaigns in the future.

To model federal campaign financing, I will first obtain the amount of money raised for the 10 most expensive Senate campaigns for every election cycle since 2008; Senate races are staggered and happen every two years, so these years would be 2008, 2010, 2012, 2014, 2016, 2018, and 2020. I will solely be looking at the top 10 campaigns in terms of money raised per cycle because 10 data points is not too many to make the graph too cluttered and obfuscating, but is also not too little for a few candidates who are able to raise abnormally high amounts to skew the data.

Next, I will plot these points on a graph and test out different trendlines to model the trend in recent years. Specifically, I will be using 3 different fits – linear, polynomial, and exponential. I will also attempt to explain and make sense of the different features of each graph. Through this investigation, I hope to learn more about the current trend in American federal campaign financing and what implications it may have for the future.

**3: Data and Basic Processing**

3.1: Raw Data

When all the raw data is plotted, the following chart is attained:

At first sight, this data seems interesting. The amount of money raised by candidates seems to be increasing over time. However, there is a dip in the 2014 and 2016 cycles, but the amount raised picks up immediately afterwards and continues to skyrocket after that.

3.2: Outlier Handling

Because I am looking at 10 of the most expensive Senate campaigns every cycle, I am bound to hit outliers; some candidates may raise an extraordinarily large which is not characteristic of the rest of the trends that year. For example, in 2010, Connecticut candidate Linda McMahon raised $50 million for her campaign, but no other candidate that cycle raised more than $30 million. Obviously, the amount that McMahon raised would skew the trend and make the amounts raised in 2010 seem higher than it actually is.

To detect and remove outliers, I will be using the interquartile range. For each year, any value that lies more than 1.5 interquartile ranges above the 3rd quartile or below the 1st quartile out of all of the values for that year will be considered an outlier.

I will discover an example outlier with the 2010 data. In 2010, the top 10 amounts of money raised for campaigns in millions of dollars was 50.29, 28.28, 23.78, 21.74, 21.41, 19.38, 18.41, 18.27, 17.76, and 17.14.

The interquartile range is the difference between the first and third quartiles. In a dataset of 10 numbers, the median would be the average between the 5th and 6th numbers, meaning the third quartile is the 3rd largest number and the first quartile is the 3rd smallest number. Hence, the interquartile range here is:

Next, any value that is more than 1.5 IQRs above the 3rd quartile or below the first quartile is considered an outlier as it deviates too much from the rest. This can be expressed in the following set of inequalities:

Here, we know that the 3rd quartile is 23.78 and the interquartile range is 5.51. Hence, the right-hand side of the first inequality evaluates to:

Any value greater than 32.045 million dollars would be an outlier. For example, the greatest value, 50.29 million dollars, is greater than 32.045 million dollars, it is considered an outlier and will be removed from the set of 2010 points. This process will be carried out with the sets of values for other years as well.

3.3: Factoring in Inflation

Next, before modeling this data, it must be first recognized that this data was collected over more than a decade – in this time, the value of the US dollar has fluctuated, due to both global trade and inflation. To gauge how much value was actually spent on campaign financing, I will adjust all dollar values to 2020 value.

An online inflation calculator will be used to find the conversions from each of the aforementioned years’ dollar values to the 2020 value.

After outliers were removed and all campaign fundraising amounts were standardized in terms of the dollar’s 2020 value, the following plot is obtained:

**4: Mathematical Modelling**

4.1: Linear Model

The first model I will try to fit is a linear model, because of its simplicity and prevalence in all sorts of contexts.

I will next derive a way to find the linear equation that best fits these points. Let the number of data points that are on the plot be . Then, each data point could be referred to as , where .

A linear line on a 2D Cartesian plane has the equation . Hence, let the line of best fit through these points be similarly defined as:

In this equation, would be the y-intercept and would be the slope of the line. This line will try to minimize the cost function , which is defined as the sum of all the distances of the points to the line squared. Mathematically, this function would be:

Because is just the definition of the line of best fit, the first definition can be substituted in for it:

To minimize the cost function , the point where with respect to and must be found. There is only one point in the function where , and that happens to be the minimum value of the cost function.

can first be differentiated with respect to . A partial derivative, denoted by the symbol instead of d, will be used because the other variables in the equation for are to be held constant in the process. Because we are trying to find when , this partial derivative will be set equal to 0:

The chain rule can first be utilized:

With some algebra, the summation can then be split up:

Next, note that as the value of is not dependent on , the summation of from 1 to is just :

Now that is clearly in the equation, algebra can be used to find :

The summation of all the and values divided by the number of points is simply the average. Note that the bar symbol denotes an average. Hence the value of that minimizes the cost function is simply:

The value of includes a reference to , so the cost function will also have to be minimized with respect to . Again, a partial derivative will be taken, and the result will be set equal to 0:

A bit more distribution can be done to make the equation simpler:

Now, because the value of is already known to be , this can be substituted into the equation to get a function of and in terms of only and :

The summation can then be split up and simplified:

Lastly, can be found through the use of algebra:

Now, the line of best fit can be calculated by substituting the values of all of the data points of amounts raised in campaigns, where the x value is the year since 2000 and the y value is the amount raised in that campaign, into these equations to find and . From that, a linear line of best fit can then be plotted.

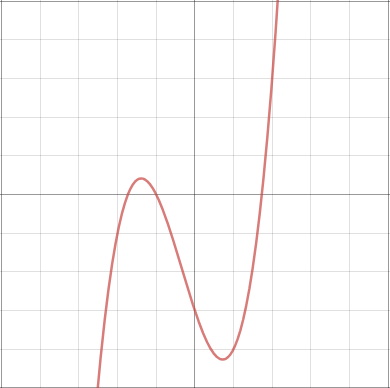
When the campaign finance data is substituted these formulas, and are obtained. From that, the following chart can be plotted:

Unfortunately, it appears that rather few of the data points are close to the line. The line significantly underestimates the amount raised in campaigns during 2008, 2010, and 2020, and significantly overestimates the amounts raised in 2014, 2016, and 2018.

Furthermore, with this model there is an x-intercept at ; this would imply that in the 2006 cycle and before, candidates raised zero or negative sums of money! This does not make logical sense, so a linear fit is probably suboptimal, and I will need to use other models.

4.2: Polynomial Model

At first sight, the next model that I thought might fit was a polynomial model. The graph that I have looks like a cubic function with a positive leading coefficient, such as the one shown below:



With a cubic function, I could obtain a model that more closely fits the actual data.

I will use Microsoft Excel to regress a cubic function through these data points. However, with the set of data points without outliers, Excel refused to regress a polynomial of power 3 through it – it insisted that the data bore a quadratic relationship, and that the leading coefficient is zero. The graph it plotted is shown below:

I was perplexed by this – the data looked like a cubic to me, but Excel rejected such a proposition. The quadratic relationship shown here is also obviously rather inaccurate – it doesn’t even pass through the ranges of the values for over half of the cycles shown. What was even more interesting was that when I tried regressing a cubic function through the data including outliers, I successfully got a cubic function. The graph that Excel plotted through the data with outliers is shown below:

I was deeply puzzled – did this mean that I should not have removed the outliers, as there would be a much more fitting relationship if they were included? Or did this simply imply that a cubic fit was simply not the best for modeling this relationship, and the cubic fit is misleading?

To begin, the cubic function seems to fit the plot with outliers included relatively well. It passes through the main clusters of data points per cycle, and manages to capture the curvature of the data.

However, it is important to note that it is unlikely that the trend of campaign fundraising dollars follows a polynomial trend perfectly. Even in nature, it is rare for a relationship between two variables to model a polynomial of a power greater than 2. Hence, having the cycle to money raised relation be modeled by a cubic function is highly improbable to begin with.

Upon further analysis of the features of this polynomial function, the model starts to make even less sense. Firstly, there is a x-intercept at ; the function is negative for any x value . This hence creates the same problem as there was in the linear model – a cubic model would mean that in cycles prior to 2008, candidates raised negative amounts of money, which does not make sense.

Furthermore, we can analyze the derivative of this function:

This derivative function has two zeros, at a bit past and a bit past . At these two points, the amount of money raised by candidates went from increasing to decreasing and from decreasing to increasing, respectively. These reversals in the trend are likely due to events that happened around these time periods. The amount of money donated to campaigns can be interpreted as having a direct relationship with the public’s interest in politics – the more the public cares about who is in power or what legislation is passed, the more money people will pour into campaigns. In 2008, the Great Recession devastated the United States economy. Hence, it is natural for political interest to have spiked after that – the American people wanted the federal government to act promptly. However, as the economy sailed to recovery in 2010, political interest would then understandably drop back down. However, there were no major events throughout the early to mid-2010s, so the rapid increase in political interest from 2014 to 2018 is unexplained. In this respect, a cubic model does not make a lot of sense either.

Therefore, even though a polynomial function may look pretty with the data locally, it is overfitting and does not accurately model the trend of Senate campaign fundraising as the trend is extrapolated. This also affirms that I was correct in removing the outliers – the outliers may make this cubic relationship even more prominent and promising-looking, but such a cubic relationship is misleading.

4.3: Exponential Model

Because of the shortfalls of both the linear model and the polynomial model, I will use one last model – the exponential model – to fit the data and see if it models the trend well. I think an exponential fit might work best because exponents and logarithms are applicable to a variety of other situations, especially with human science. For example, in economics, ideas such as interest or inflation are based on exponential functions. Other concepts such as population growth are also often exponential.

When using Excel to regress the data exponentially, the following chart is obtained:

This plot looks more legitimate than the linear and polynomial fits. To begin, this function fits the data better than the linear plot – there is no significant overfitting or underfitting and the line seems to be able to mostly keep up with the data points. Upon first sight, even though this function seems to fit the points less well than the polynomial model, it still captures the shape of the data. Nevertheless, social science cannot be quantified by pure math, so deviation from the trendline is expected to some extent.

More importantly, the physical features of this model are more reasonable. There is a horizontal asymptote at as , which makes sense. It means that at elections further back in the past, less money was raised for campaigns.

Furthermore, the derivative of this function makes more sense as well. When differentiated using the chain rule, the following is obtained:

The first derivative is another exponential function with all positive coefficients, meaning that it is always positive. This means that campaign fundraising dollars are expected to keep increasing as time passes. It is also interesting to note that the first derivative is just 0.1028 times the original function – this means that the first derivative is increasing as well, meaning that the amount of money being raised by Senate candidates is increasing at an increasing rate year after year. This is consistent with what many news sources say about federal elections becoming more and more expensive as time goes on, so such an exponential model makes logical sense.

One deficiency with this model is the fact that it severely underestimates the set of values in 2020, by a margin as wide as the range of the values for some other years. However, this and other deviations can be explained by special circumstances in those years. For example, in 2020, amid the coronavirus and other social issues that have risen to the forefront of the political agenda, the public has become more attentive than ever to politics. Candidates have hence been able to raise at all-time highs because there is great public interest. Similarly, the trendline underestimating the amount raised in Senate campaigns in 2008 and 2010 can be possibly attributed the Great Recession, while the trendline overestimating the amount raised in 2014, 2016, and 2018 can be explained by the relative lack of major political, social, or economical events in those years.

In any case, social science is predictably irrational, and fundraising trends cannot be characterized perfectly by mathematical models. Overestimating and underestimating from time to time is thus expected. With everything that was mentioned in consideration, an exponential model seems to best model the trend of the money raised by Senate race candidates over time, both from data-fitting and logical standpoints.

**5: Conclusion**

5.1: Conclusion and Implications

In conclusion, the amount of money that Senate candidates are raising for their campaigns is roughly increasing every cycle. Specifically, the amount raised increases in an approximately exponential manner as time passes. An exponential relationship was established because it was the best of the 3 models that I tried fitting – linear, polynomial, and exponential – mainly for the following reasons:

* It fit and captured the shape of the data reasonably well
* The features of the graph (intercepts, asymptotes, etc.) made logical sense
* The derivative of the function makes sense as well and is consistent with what is often said about federal campaign costs
* Deviations from the trendline can be explained

Because Senate races are becoming more and more expensive, it can be reasonably well inferenced that House and presidential races are likely costing more and more money as well. If this trend continues, it means that much more money will be poured into politics as time goes on, which is detrimental to the US economy and consistent with what critics would deem as “corruption” in the government. If the nation was aware of this, would we perhaps be led to reevaluate our system and our choices?

5.2: Evaluation and Extensions

This paper had both strengths and weaknesses. One strength was that I tested out different models and justified why each model would fit. I also only used data that was likely representative – I took out outliers and also adjusted values to deflation.

However, a weakness of this paper is that I only tried to fit 3 models. I attempted to choose models that would make sense, but the linear and polynomial models did not capture the essence of the data, while the exponential model may have seemed underfitting due to deviation, especially in 2020. Furthermore, my original objective was to track trends in federal campaign financing. Tracking the top 10 amounts of money raised for Senate races for each 2-year cycle can help demonstrate the general trend of campaign financing, but it is not the whole story and can arguably be misrepresentative of federal campaign financing as a whole. It may be the case that my data is drastically different from trends in campaigns for the House of Representatives, or even from Senators who raised lesser amounts of money each year.

If I had more time, I would definitely collect and process greater amounts of data to get a more complete picture of campaign financing. For sake of clarity and avoiding cluttered data, I only chose the top 10 amounts of money raised by Senate candidates each cycle, but I could possibly use data from all Senators next time. I could also analyze campaigns in the House of Representatives or presidency, to see if the trends there are consistent with my findings here. Furthermore, I could try out some other models, such as a power function, to see if it fits my data better. Lastly, I may even specifically compare campaign financing before and after monumental events such as 911, the 2002 McCain-Feingold Act, the Great Recession, etc., to see how they shift the amount of money injected into the intricate system of campaign financing.

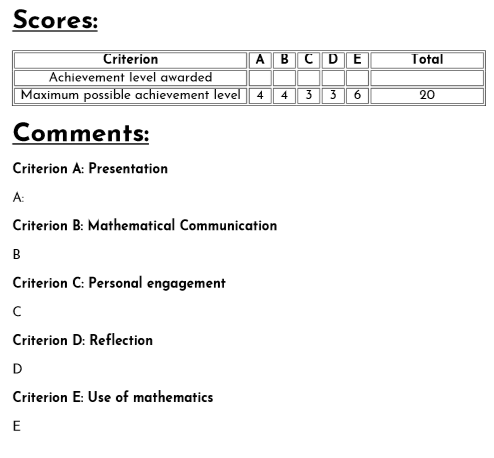
**6: Bibliography**

[this is a first draft so I’m putting links for now; I might have more sources later, and I like converting to MLA/APA all at once in the end]

* <https://www.mtsu.edu/first-amendment/article/1078/federal-election-campaign-act-of-1971>
* <https://www.opensecrets.org/elections-overview/fundraising-totals>
* <https://www.statisticshowto.com/find-outliers/>
* <http://web.mnstate.edu/peil/MDEV102/U4/S36/S363.html>
* <https://towardsdatascience.com/linear-regression-derivation-d362ea3884c2>
* <https://medium.com/@lachlanmiller_52885/understanding-and-calculating-the-cost-function-for-linear-regression-39b8a3519fcb>

Note to self: my data is sparse right now and I want data for more years. Right now opensecrets.org displays an “there was an error” message when I try to access years from 2006 and prior; I may be able to get more data if this is fixed soon. I may also switch to graph campaign spending instead of campaign fundraising

**FIRST DRAFT ONLY: self read-and-score**

****

Criteria A: 2/4 – I think my work was decently well organized and made coherent sense. I may have included some things that were extraneous, but overall I probably structured everything in a standard and understandable way

Criteria B: 2/4 – The math was expressed clearly and I used symbols appropriately. I’m not entirely sure I did everything as I was supposed to, but I think my notations/symbols/terminology are acceptable throughout

Criteria C: 2/3 – I think my topic is decently unique and I can get personal engagement points for that. I also try to demonstrate that I’m interested in it and I’m knowledgeable about the topic, proving that I didn’t just happen to find this topic while google searching for data to model for my IA. I also explain why I used the math that I did

Criteria D: 1/3 – My reflection is severely lacking; I couldn’t think of many good things to say or many strengths/extensions that I could have

Criteria E: 4/6 – I used a variety of topics, ranging from statistical analysis to partial derivatives to interpreting features of functions. This demonstrates decent use of what we learned in class, but I don’t think I did exemplary on this

Total: 11/20