George Daher
STAT 277 – Project Written Introduction & Analytical Plan
11/4/24

Written Introduction

A very large part of the democratic election process in the United States is fundraising. This can happen in a number of different ways depending on a candidate in their policy positions. For example, some politicians rely on a large number of small donations from individuals, whereas other politicians may rely on large donations from specific individuals, companies or political interest groups. This data for these donations is largely public, as election spending laws require a significant amount of transparency. As a result, much of the data on political fundraising and spending with relation to election campaigns, such as federal election campaigns for congress or the presidency, is public available on sites like fec.gov, the Federal Election Committee Website, or publicreporting.elections.ny.gov, the New York State campaign finance website. Using this data, there is a significant amount of research on the effect of campaign fundraising and expenditure on the outcome of elections.

My goal is also to understand the relationship between election spending and expenditure with election outcome, but I want to bring other ideas to understand this relationship more deeply. News articles, make the claim that a majority of the time the better funded candidate wins (Koerth 2018). However, addressing confounding variables has led others to conclude that election spending has a less severe impact on election outcome (Levitt 1994). Investigating this claim is intended to be the very beginning of my research and analysis process. I hope to use this to investigate deeper questions between the outcomes of elections on the fundraising/expenditure process. One of the main things I aim to consider is whether the type of fundraising has an impact on election odds, there are only a few research articles that consider this, and the ones that do typically focus on a small subset, such as open seat house races (Alexander 2005), rather than a more general overview across the country than I intend to do. Additionally, further distinctions could be made about fundraising and expenditures than just categories and types. For example, prior research has found that incumbent candidates that spend more money are more likely to lose their elections (Ansolabehere et Gerber 1994). There is clearly no causation, but the correlation is there, so it is necessary to consider in a larger analysis.

In general, there is a lot of statistical research done in political science. This means there is a large amount of research addressing a variety of facets of the election process, ranging from that related to campaign fundraising and expenditures, all the way to psychological considerations of voters themselves. This saturation could make an initial investigation into the field daunting, especially as someone who is not knowledgeable about political science, despite an interest in politics. However, this saturation of research is a benefit, especially with my goal of a more general overview and investigation of the relationship of election outcomes with fundraising and expenditure. Each paper in the field can offer a potential sub-category of interest, such as the consideration categorizing races into ones with open seats or incumbents (as referenced in the journal article by Alexander, 2005 mentioned before). Additionally, I am interested in looking at some of the more advanced statistical techniques used in election modeling, such as adding a spatial element for predicting election outcome or considering donations or expenditures. There are also a number of prior papers in the field that address this. I may also consider statistical techniques that I am familiar with myself, such as Bayesian modeling techniques. The large number of prior articles offer both general guidance for research I can do as well as ideas to consider in an understanding of the data as a whole.

With all of this considered my current research question, although unlikely to be in its final form, is as follows: Do characteristics of campaign have a relationship with election outcome in federal elections for the House of Representatives over 3 election cycles (2014, 2016, and 2018)? I believe that there is a relationship and that characteristics of the funding have a strong relationship with election outcome, on top of the relationship between total money raised and election outcome.

The main response will be a binary variable of whether an election was won or not. Then our independent variables will be characteristics of fund-raising including total money raised, with information about what sources this money has come from, where the money has been donated from (spatial element), whether the money was donated by individuals who often donate to political campaigns, and a number of other potential considerations. The data for all of this will mainly be acquired from the Federal Election Commission website, but other auxiliary data sources such as the US census may be used to serve as structural variables to potentially explain variations within the dataset. As I investigate my research question, as well as potentially tweaking it along the way, I may also consider incorporating other elements of election models and political science into my research.

Bibliography

Alexander, B. (2005). Good Money and Bad Money: Do Funding Sources Affect Electoral Outcomes? *Political Research Quarterly*, *58*(2), 353–358. https://doi.org/10.2307/3595635

Ansolabehere, S., & Gerber, A. (1994). The Mismeasure of Campaign Spending: Evidence from the 1990 U.S. House Elections. *The Journal of Politics*, *56*(4), 1106–1118. https://doi.org/10.2307/2132077

Briffault, R. (1999). Public Funding and Democratic Elections. *University of Pennsylvania Law Review*, *148*(2), 563–590. https://doi.org/10.2307/3312798

Gerber, A. (1998). Estimating the Effect of Campaign Spending on Senate Election Outcomes Using Instrumental Variables. *The American Political Science Review*, *92*(2), 401–411. https://doi.org/10.2307/2585672

Jacobson, G. C. (1990). The Effects of Campaign Spending in House Elections: New Evidence for Old Arguments. *American Journal of Political Science*, *34*(2), 334–362. https://doi.org/10.2307/2111450

Klumpp, T., Mialon, H. M., & Williams, M. A. (2015). Leveling the Playing Field? The Role of Public Campaign Funding in Elections. *American Law and Economics Review*, *17*(2), 361–408. http://www.jstor.org/stable/24735755

Koerth, M. (2018, September 10). *How Money affects elections*. FiveThirtyEight. https://fivethirtyeight.com/features/money-and-elections-a-complicated-love-story/

Levitt, S. D. (1994). Using Repeat Challengers to Estimate the Effect of Campaign Spending on Election Outcomes in the U.S. House. *Journal of Political Economy*, *102*(4), 777–798. http://www.jstor.org/stable/2138764

Schuster, S. S. (2020). Does campaign spending affect election outcomes? New evidence from transaction-level disbursement data. *The Journal of Politics*, *708646*, 000–000. https://doi.org/10.1086/708646

Stratmann, T. (2006). Contribution Limits and the Effectiveness of Campaign Spending. *Public Choice*, *129*(3/4), 461–474. http://www.jstor.org/stable/25487608

Analytical Plan

The question I am studying is whether the relative amount of money raised in a congressional election impacts the result, while accounting for confounding variables such as population, housing and economic information.

My hypothesis is that we can create a statistically significant logistic model to predict whether an election will be won or not based on the campaign finance information (and the confounding variables). More specifically, the null hypothesis is that all variables in the logistic model are not significant, depending on how the model is viewed these coefficients may all be 0, or all 1 (odds or odds ratio) in a null hypothesis.

I am using three different data sources: campaign finance information and election results information from the Federal Election, and 5 year estimates from 2014 for demographic data collected from the American Community Survey. The campaign finance dataset includes the main explanatory variables of amount of money collected, and from which sources this money was collected. The election results dataset will be used to determine whether an election was won or not, which will be displayed as a binary response variable for the logistic model used. Finally, the American Community Survey information will be used as confounding variables to incorporate in the model, as I anticipate there will be some relationship in according to district income, educational demographic or something similar, with campaign finances raised and potentially the effect of campaign finances on an election result.

The first data source is the Federal Election Commission, for two different datasets. The first dataset is summaries of campaign finance information about candidates for congressional races. This data is collected to ensure that campaign finance law is followed with the goal of fair, free and transparent elections. The data is separated by year into different datafiles for each federal election cycles, which is every 2 years, and 2014, 2016 and 2018 are considered in the scope of this project. Each data point in these files is a campaign run by an individual for federal office, there are 1841 observations in the 2014 files, and 1897 observations in each of the 2016 and 2018 files. For each of these datasets there are 30 variables recorded per observation. This dataset contains the main explanatory variables of campaign finance amounts collected (and basic information about the sources from which it was collected).

The data dictionary is as follows:

Column name	**Field name**	**Position**	**Null**	**Data type**	**Description**	**Example data**
CAND_ID	Candidate identification	1	N	VARCHAR2 (9)		H8VA01233

CAND_NAME	Candidate name	2	Y	VARCHAR2(200)		Martha Washington
CAND_ICI	Incumbent challenger status	3	Y	VARCHAR2(1)		I
PTY_CD	Party code	4	Y	VARCHAR2(1)		NON
CAND_PTY_AFFILIATION	Party affiliation	5	Y	VARCHAR2(3)		NON
TTL_RECEIPTS	Total receipts	6	Y	Number(14,2)		345,456.34
TRANS_FROM_AUTH	Transfers from authorized committees	7	Y	Number(14,2)		4000.00
TTL_DISB	Total disbursements	8	Y	Number(14,2)		175645.21
TRANS_TO_AUTH	Transfers to authorized committees	9	Y	Number(14,2)		0.00
COH_BOP	Beginning cash	10	Y	Number(14,2)		845901.23
COH_COP	Ending cash	11	Y	Number(14,2)		915671.43
CAND_CONTRIB	Contributions from candidate	12	Y	Number(14,2)		500.00
CAND_LOANS	Loans from candidate	13	Y	Number(14,2)		250000.00
OTHER_LOANS	Other loans	14	Y	Number(14,2)		0.00
CAND_LOAN_REPAY	Candidate loan repayments	15	Y	Number(14,2)		100000.00
OTHER_LOAN_REPAY	Other loan repayments	16	Y	Number(14,2)		0.00
DEBTS_OWED_BY	Debts owed by	17	Y	Number(14,2)		250.00
TTL_INDIV_CONTRIB	Total individual contributions	18	Y	Number(14,2)		450000.00
CAND_OFFICE_ST	Candidate state	19	Y	VARCHAR2(2		VA
CAND_OFFICE_DISTRICT	Candidate district	20	Y	VARCHAR2(2)		01
SPEC_ELECTION	Special election status	21	Y	VARCHAR2(1)	Election result data included in 1996-	W

					2006 files only.	
PRIM_ELECTION	Primary election status	22	Y	VARCHAR2(1)	Election result data included in 1996-2006 files only.	L
RUN_ELECTION	Runoff election status	23	Y	VARCHAR2(1)	Election result data included in 1996-2006 files only.	W
GEN_ELECTION	General election status	24	Y	VARCHAR2(1)	Election result data included in 1996-2006 files only.	L
GEN_ELECTION_PRECENT	General election percentage	25	Y	Number(7,4)	Election result data included in 1996-2006 files only.	63.2
OTHER_POL_CMTE_CONTRIB	Contributions from other political committees	26	Y	Number(14,2)		200000.00
POL_PTY_CONTRIB	Contributions from party committees	27	Y	Number(14,2)		200000.00
CVG_END_DT	Coverage end date	28	Y	DATE(MM/DD/YYYY)	Through date	10/19/2018
INDIV_REFUNDS	Refunds to individuals	29	Y	Number(14,2)		4000.00
CMTE_REFUNDS	Refunds to committees	30	Y	Number(14,2)		100

The second dataset is also from the Federal Election Commission. It is a summary of the results (votes cast and for which candidate) of the various federal elections that occurred during a particular election cycle. This data is collected as part of the election process. The data is separated by year into different datafiles for each federal election cycles, which is every 2 years, and 2014, 2016 and 2018 are considered in the scope of this project. Within these datafiles (which are Excel files) there a sheets for presidential (if applicable), Each data point in these files is an individual who received votes during a federal election during an election cycle. This data requires a significant of cleaning and reformatting. The data is partially wide and long, as there are rows of the dataset that exist only as partial totals for some subset of the dataset. Additionally, this table contains primary and general election candidates and only certain candidates are relevant in the general election results (those that one the primary). There are multiple sheets of relevant data and the number of rows varies per election cycle, and by house or senate results. There are about 500-600 rows in the senate election results (across all 3 cycles) and about 4000 results rows in the house of representative election results. For each election cycle (for both house and senate results), there are 16 variables observed.

The data dictionary is as follows:

Variable Number	Variable Name	Description
1	STATE ABBREVIATION	
2	STATE	
3	(I)	Indicates whether a candidates if incumbent
4	FEC ID#	
5	(I)	
6	CANDIDATE NAME (First)	
7	CANDIDATE NAME (Last)	
8	CANDIDATE NAME	
9	TOTAL VOTES	Indicates whether a data row is a total row rather than an observation
10	PARTY	
11	PRIMARY VOTES	
12	PRIMARY %	
13	RUNOFF VOTES	
14	RUNOFF %	
15	GENERAL VOTES	
16	GENERAL %	

The third datasets are from the American community survey. There are three separate datasets for economic, demographic and social information separated by Congressional District. These datasets contain a number of data columns with 5 year estimates about various statistics concerning a congressional district. Examples of some statistics included are median income, total population (broken down by race) and education status obtained. For each of these statistics there is an estimate, a estimate margin of error, a percentage of the population the statistic entails (if applicable), and a percentage margin of error. Each of these datasets has 435 rows, one observation for each of the congressional districts being considered, and each of these rows has a large number of variables (greater than 500) that are the economic, demographic and social statistics being considered.

The data dictionary for these datasets is omitted because of the size. During the EDA, specific statistics will be parsed from this dataset and considered in analysis.

The EDA is especially important because there are large number of variables to potentially include in the logistic model being created. By looking at singular variables, or subsets of variables and their relationship with the response, we can get a decent idea of whether a variable should be considered in the model. Additionally, by waiting until creating the model itself to collapse the election results to binary (win or no win), and keeping them instead as a percentage of votes received, then the EDA can be more effective.

The significance of the resulting model will be evaluated as a whole in order to answer the research question. Additionally, Wald tests can be used for individual variable significance in the model. Depending on the strength of the results, cross-validating the model could also be done to test its strength and validity.

The assumptions for logistic regression are fairly weak compared to those of linear regression. The only required assumptions for the data are that the response is binary, the observations are independent, and that there is minimal multicollinearity. The binary response variable can be achieved by collapsing election results to a win or a loss. The independence of the observations is not fully given, but the large sample size (across many election races and multiple election cycles) can mitigate this. Additionally treating the data as a matched pairs due to the fact that the elections are between two individuals, would remove some of the problems with independence among the data. Some elements of this independence cannot be fully remedied but this lack of independence would also persist across future elections so the results remain meaningful despite this. Finally, the multicollinearity of the data can be examined and tested for. All the assumptions can be checked for the logistic regression technique used.

Data Cleaning/Final Data Dictionary

A significant amount of work was required in order to clean the data and make it usable for analysis. As specified in the analytical plan, there were a lot of variables to consider in the datasets used. A lot of these variables ended up being dropped because they did not hold valuable information. Other than this, the formatting of some of the datasets was problematic for any analysis, and required extensive reformatting as well as proper NA data value formatting or dropping. Additionally, there was the added fact that a significant amount of the data has a matched pairs structure. In most of the elections, two individuals faced against each other, and accounting for this structure while also keeping the observations separate was another consideration in reformatting.

After extensive reformatting, outline in a data cleaning R markdown file, the final data dictionary was as follows:

Data Name	**Description (if needed)**	**Format**
STATE_ABBREVIATION		2 Character
D	District Number	Discrete Numeric
PARTY	Political Party	1 Character
GENERAL_VOTES	Votes received in General Election	Numeric
GENERAL_PROP	Proportion of votes received in General Election	Numeric: 0 to 1
GE_WINNER_INDICATOR	1 if general election won, 0 if general election lost	Boolean: 0 or 1
Incumbent_challenger_status	C if challenger candidate, I if Incumbent candidate, O if open seat,	1 Character
Total_receipts		Numeric
Transfers_from_authorized_commi	Transfers from authorized committees	Numeric
Total_disbursements		Numeric
Transfers_to_authorized_committ	Transfers to authorized committees	Numeric
Beginning_cash		Numeric
Ending_cash		Numeric
Contributions_from_candidate		Numeric
Loans_from_candidate		Numeric
Other_loans		Numeric
Candidate_loan_repayments		Numeric
Other_loan_repayments		Numeric
Debts_owed_by		Numeric

Total_individual_contributions		Numeric
Contributions_from_other_politi	Contributions from other political committees	Numeric
Contributions_from_party_commit	Contributions from party committees	Numeric
Refunds_to_individuals		Numeric
Refunds_to_committees		Numeric
Year	Current Year – 2000 (i.e 2014 is 14, etc)	Numeric

In the analysis portion, some of these variables are transformed or collapsed. For example, all third parties (non democrat/republican) are collapsed to a singular third party indicator. Another instance is converting the financial variables to proportions compared to the total race in a particular district. This would note what proportion of the spending/receiving in a particular state and district was done by a specific candidate. Additionally, some of the variables are log transformed in the analysis as they are financial data points, and such data is often log transformed in order to achieve some normality while preserving ordering and structure.

Exploratory Data Analysis – EDA

NOTE: All EDA table/plots are in a separate pdf due the result length

Due to the number of variables in the dataset, there are a lot of relationships to consider within the EDA. We can begin with the single variable analysis, which is summary statistics for the numeric variables, and frequency tables for the numeric variables.

The first thing to note is that there is very little missingness in the data. The only variable with missing values is incumbent challenger status, and in that case, there are still only 3 missing values. This is due to the fact that a lot of the missingness in the data was handled during the data cleaning step. In the EDA, we can also note that the distributions of D, district number, and state abbreviation are not important to us. They follow a known pattern, the 435 house districts that make up the US, and we only include them in the data because of an interaction effect that may be present in the data.

We first consider the distributions of the 3 potential response variables, a Boolean variable for whether an election was won or not, the number of votes a candidate received in the general election, and the proportion of votes they received in the general election. The purpose of keeping all three of these response variables at this stage is to see how they differ as response variables. Each of them holds a different value as a response variable and we expect slightly different results, although we know by construction that these three variables will be significantly correlated. Due to the number of columns in the data set (17 to be precise), we consider the relationship between all the other variables and each of these three response variables. Then by looking at a correlation table, we can examine relationships between correlated variables. On their own, these response variables are somewhat symmetrically distributed and unimodal, which generally makes sense considering the balanced nature of elections and the way congressional districts are constructed.

We can get a brief summary of the character variables of political party and incumbent challenger status by looking at frequency tables. For political parties, we see there are roughly even a number of Democrats and Republicans, and there are a smaller but still significant number of third-party candidates. For incumbent challenger status, we see that there are a similar number of incumbent candidates and challenger candidates, and a smaller but still significant number of candidates running for an open congressional seat.

The financial variables are all very right-skewed, meaning the distributions are saturated on the lower end but have a significant amount of higher values and high outliers. This makes sense in the context of financial variables that represent some dollar amount, as this is often the case. There are a few options in this case. The first option is that by isolating each race, we can convert each of these variables to a proportion, where we can view the total amount of money received, for example, as a proportion of the total money received by all candidates in a certain district race. The second option, which is fairly common for analyses involving financial variables (often salary), is to log transform all these variables in order to achieve some level of normality. The problem with this approach is that there are a large number of 0 values for some of the financial variable columns, but we can address this by adding some small constant to all values in order to account for this. Looking at the log transformations for all the financial variables, we see some level of normality, although we lose the comparative structure of comparing the amount spent between two candidates in a district. We use these log-transformed values for the rest of the EDA and analysis (except for the correlation table).

The first response variable we compare the explanatory financial variables to is the Boolean election win/loss value. For each explanatory variable, we can observe side-by-side boxplots of the financial variables grouped by whether the election was won or lost. In these, we see that most of the variables seem to have a different distribution depending on whether an election was won or lost, and for some, the relationships seem negligibly different. From this observing scatter plots between the financial variables and votes received and the proportion of votes received, a relationship is less clear. A number of the plots seem to show some linear relationship but others show no clear relationship. On its own, this is not very meaningful, but it indicates that some of the variables may not be important in an eventual model that is created.

The final step of the EDA is a correlation table. This displays the correlations between all the numeric variables in our dataset, which creates a large table (20 by 20) because there are many numeric variables in the dataset. Notably, most of the correlations in the table are fairly low, except for those between the response variables, as we would expect. Additionally, many of the contribution variables are related to the total receipts, which is also understandable, as they contribute directly to the larger sum variable. The main conclusion that we can reach from this correlation table is that it is unlikely that multicollinearity will be a problem in any models that we create, but it should not be ignored entirely.

Overall, the EDA does not show any very specific results about the data, but we can get a better understanding of the distributions of the variables and their relationships. Additionally, the relationships between the explanatory and response variables hint that further analysis and statistical testing are worth doing, as there are no clear results that answer the research question.