Q3 Federal Elections Exercise

Unit 9 Coding Homework - W203 Section 8

Chase Madson

Contents

The Setup: There is a lot of money that is spent in politics in Presidential election years. Like, a lot, a lot. Estimates and analysis from the US Federal Election Comission, puts the total amount at about $14,400,000,000 ($14.4 billion USD). For context, Twitter's 2020 annual revenue was about $3,500,000,000 ($3.5 billion USD).

The Work:

The package `fec16` is a compendium of spending and results from the 2016 election cycle. In this dataset are 9 different datasets that cover:

- candidates: candidate attributes, like their name, a unique id of the candidate, the election year under consideration, the office they're running for, etc.
- results_house: race attributes, like the name of the candidates running in the election, a unique id of the candidate, the number of general_votes garnered by each candidate, and other information.
- campaigns: financial information for each house & senate campaign. This includes a unique candidate id, the total receipts (how much came in the doors), and total disbursements (the total spent by the campaign), the total contributed by party central committees, and other information.

Your Task: *Describe the relationship between spending on a candidate's behalf and the votes they receive.*

If it is helpful to structure your response, you might want to place yourself into a scenario where you are advising a person or business about whether they should make a political donation. While the benefits that accrue as a result of a successful investment are unclear, you can be quite sure that investing with no return (i.e. more spending does not increase the chances of winning) is a bad idea.

Your Work:

- We want to keep this work *relatively* constrained, which is why we're providing you with data through the `fec16` package. It is possible to gather all the information from current FEC reports, but it would require you to make a series of API calls that would pull us away from the core modeling tasks that we want you to focus on instead.

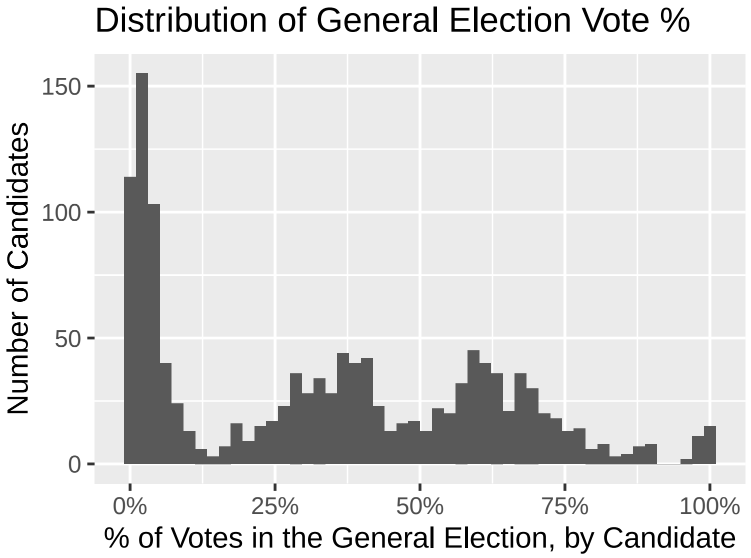
- Throughout this assignment, limit yourself to functions that are within the `tidyverse` family of packages: `dplyr`, `ggplot`, `patchwork`, and `magrittr` for wrangling and exploration and `base`, `stats`, `sandwich` and `lmtest` for modeling and testing. You do not *have* to use these packages; but try to limit yourself to using only these.

- Our choice to encourage you to use only these packages is to try to cut down on the amount of searching that you do: to help you avoid looking for the *"one package that does the thing I need it to do."* Certainly, such a package exists, but it will very likely be more productive for you to write things yourself than to try and find it for this homework.

Exploring

3.1 Make Histograms pt 1

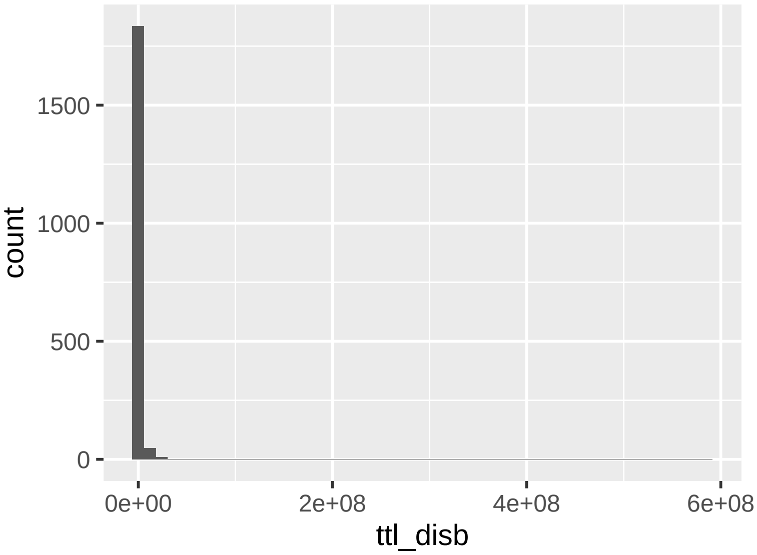
In separate histograms, show both the distribution of votes (measured in results_house$general_percent for now) and spending (measured in ttl_disb). **How is the distribution of general_percent shaped?**



This distribution appears to skew to the right

3.2 Make Histograms pt 2

In separate histograms, show both the distribution of votes (measured in results_house$general_percent for now) and spending (measured in ttl_disb). **How is the distribution of ttl_disb shaped?**



3.3 Build a Data Frame pt 1

Create a new dataframe by joining results_house and campaigns using the inner_join function from dplyr on the field cand_id. **How many rows are in the new dataframe?**

```
## [1] 1342
```

> After joining the two tables we have 1342 rows

3.4 Build a Data Frame pt 2

Create a new dataframe by joining results_house and campaigns using the inner_join function from dplyr on the field cand_id. **How many columns are in the new dataframe?**

```
## [1] 37
```

> After joining the two tables we have 37 columns

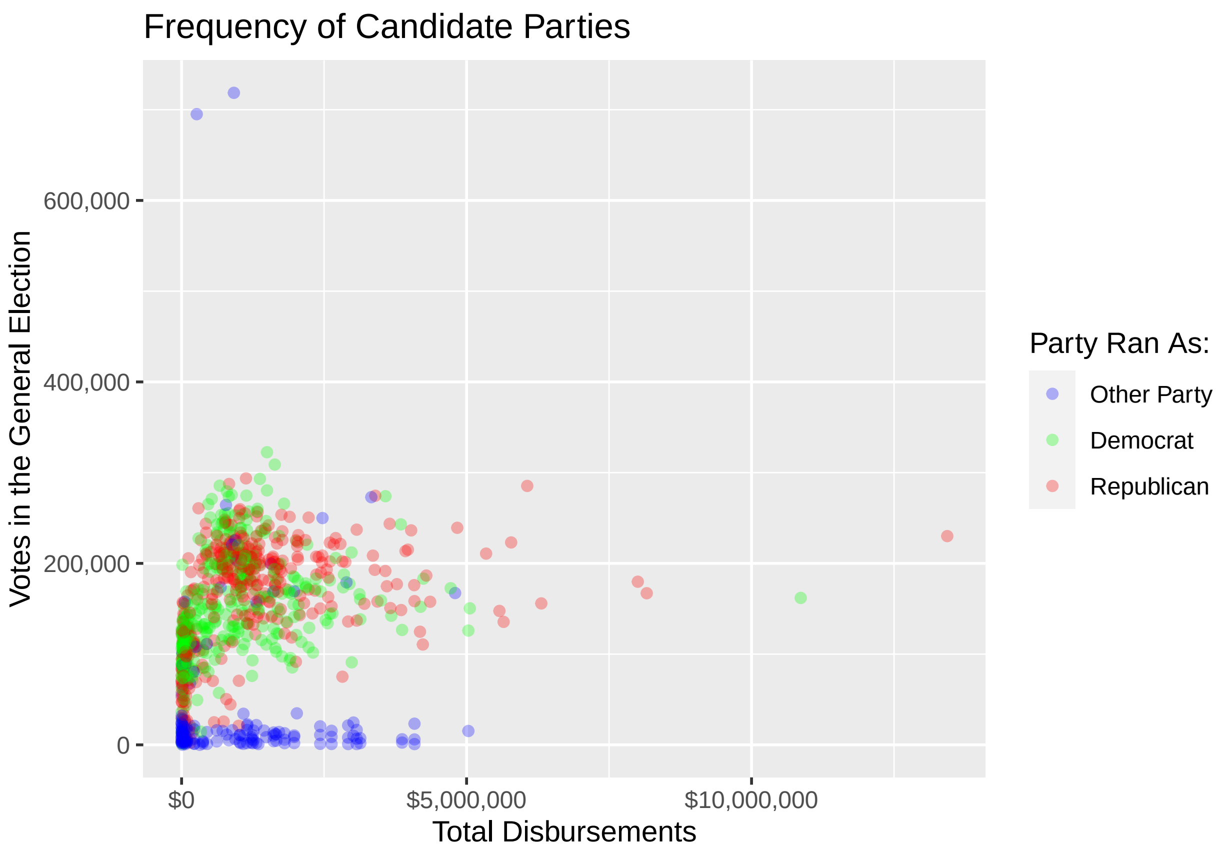
3.5 Indicator Variables and Scatter Plot

Create a new variable candidate_party to indicate whether each individual is a "Democrat", "Republican" or "Other Party" based on the party column

Note: Both the party column from results_house table and the cand_pty_affiliation column from the campaigns table contain information about the candidate's party. party is the party that the candidate actually ran as in their race, while cand_pty_affiliation is the party that the candidate self-identified as with the FEC. The two columns may not necessarily match. (e.g. Let's think about Bernie Sanders. He has identified himself to the FEC as an independent. When he ran for president, he actually ran as a democrat.)

Use ggplot to produce a scatter plot of general_votes on the y-axis and ttl_disb on the x-axis and color the points by party membership. You do not need to apply any transformations to the variables before plotting.

```
## Warning: Removed 462 rows containing missing values (geom_point).
```



```
## Warning: Removed 462 rows containing missing values (geom_point).
```

Regression

3.6 Evaluate large sample assumptions

Produce a linear regression with the outcome general_votes on ttl_disb and candidate_party. Do not apply transformations to these variables.

Evaluate the large-sample linear model assumptions presenting evidence based on your background knowledge, visualizations, and numerical summaries. Please limit your answer for this subpart evaluating both assumptions to no more than a page in total. Written responses should be no longer than 5 sentences for each assumption.

Upload your argument about the large-sample assumptions here. Your answer for this part cannot exceed a page

```
##
## Call:
## lm(formula = general_votes ~ ttl_disb + candidate_party, data = .)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -146242  -38135  -11551   37488  679443
##
```

```
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)    
## (Intercept)               2.698e+04  5.445e+03   4.954 8.71e-07 ***
## ttl_disb                  1.326e-02  1.742e-03   7.614 6.90e-14 ***
## candidate_partyDemocrat   1.131e+05  6.154e+03  18.378  < 2e-16 ***
## candidate_partyRepublican 1.196e+05  6.191e+03  19.325  < 2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
## 
## Residual standard error: 64490 on 876 degrees of freedom
##   (462 observations deleted due to missingness)
## Multiple R-squared:  0.3589, Adjusted R-squared:  0.3567 
## F-statistic: 163.5 on 3 and 876 DF,  p-value: < 2.2e-16
```

For OLS in a large sample ($n > 100$), we need to verify a few assumptions.

1. The data must be IID

    - Each observation is drawn from political candidates runnning in a 2016 election. We may see violations of IID in this circumstance given that there is strong regional clustering of a candidate's success based on their party and geographic location. If there are two candidates from the same state, there is likely some covariance between their political parties. Therefore we would consider this assumption likely violated.

2. A unique BLP exists

    - The existence of a BLP may be in question since the predictor ttl_disb, which we saw before in **Q3.2** was severely right-skewed. This causes the covariance between this predictor and the other variables to be infinite. Therefore we would consider this assumption likely violated, and log-transformation may be an appropriate remedy. However, if there is a lot of skew in the data, then the n must be large enough to overcome it. Here we have $n > 1000$ in addition to a strong skew, which may be enough.
    - The uniqueness of the BLP appears to be valid - it does not appear reasonable to write one predictor as a linear combination of another. The budget of a campaign may have some relationship to the party (especially for third-party candidates), but this is not extreme enough to invite perffect collinearity.

For OLS in a large sample ($n > 100$), we need to verify a few assumptions.

3.7 Build a stargazer table

Produce a latex formatted Stargazer table that shows the model that you created. Make sure that you are using appropriate standard errors. This table should be of the same quality that you would upload for a lab (title, english words instead of variable_names, etc.)

Upload the table along with the stargazer command used to generate it. You should also include any commands used to generate your appropriate standard errors.

3.8 Money's Relationship with Votes

3.9 Party's Relationship with Votes

Table 1:

	Dependent variable:
	general_votes
ttl_disb	0.013^{***}
	(0.002)
candidate_partyDemocrat	$113,100.600^{***}$
	(6,154.125)
candidate_partyRepublican	$119,634.600^{***}$
	(6,190.662)
Constant	$26,975.280^{***}$
	(5,444.769)
Observations	880
R^2	0.359
Adjusted R^2	0.357
Residual Std. Error	64,486.840 (df = 876)
F Statistic	163.499^{***} (df = 3; 876)
Note:	*p<0.1; **p<0.05; ***p<0.01