Politics Are Afoot!

w203: Statistics for Data Science

The Setup

There is *a lot* of money that is spent in politics in Presidential election years. So far, estimates have the number at about $11,000,000,000 (11 billion USD). For context, in 2019 Twitter's annual revenue was about $3,500,000,000 (3.5 billion USD).

The work

Install the package, `fec16`.

```
## install.packages('fec16')
```

This package 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.

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.

```r
library(tidyverse)
```

```
## Warning in system("timedatectl", intern = TRUE): running command 'timedatectl'
## had status 1
```

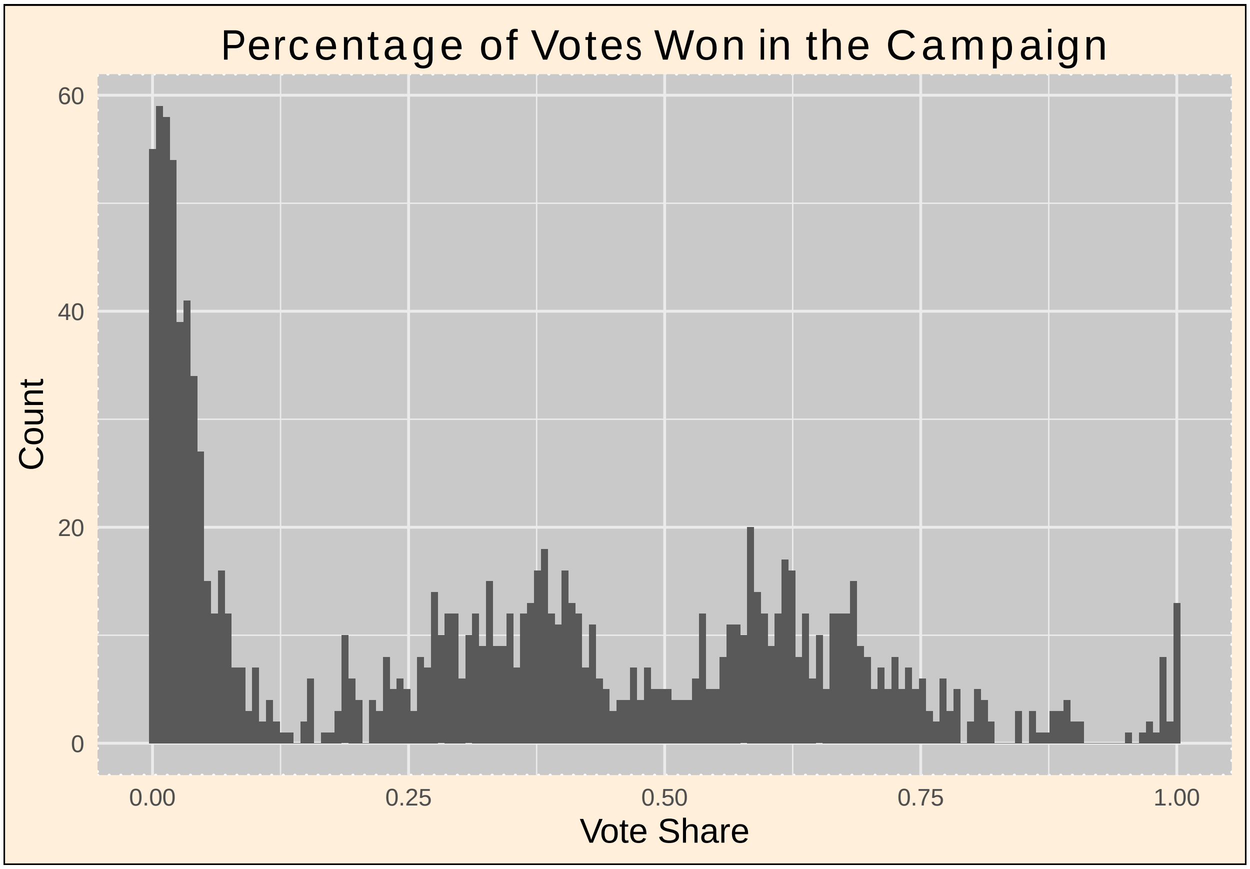
```r
library(magrittr)
library(ggplot2)
library(patchwork)
library(sandwich)
library(lmtest)
library(fec16)
theme_set(theme_minimal())
knitr::opts_chunk$set(dpi = 300)
```

```r
candidates    <- fec16::candidates
results_house <- fec16::results_house
campaigns     <- fec16::campaigns
```

1. What does the distribution of votes and of spending look like?

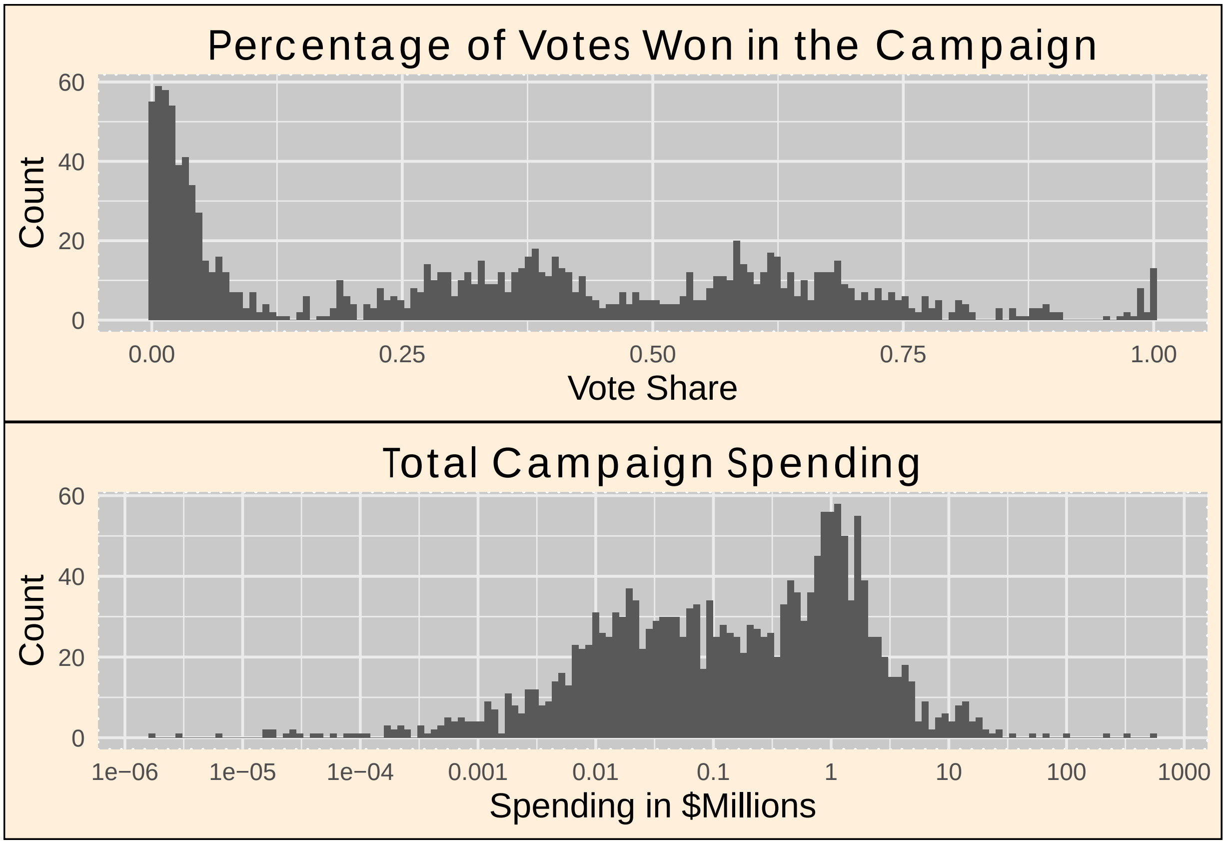
("ggThemeAssist") 1. (3 points) In separate histograms, show both the distribution of votes (measured in `results_house$general_percent` for now) and spending (measured in `ttl_disb`). Use a log transform if appropriate for each visualization. How would you describe what you see in these two plots?

```r
library("ggThemeAssist")
general_percent_histogram <- ggplot(data=subset(results_house, !is.na(general_percent)), aes(x=general_
geom_histogram(bins = 150) +
  labs(
    title = 'Percentage of Votes Won in the Campaign',
    x = 'Vote Share', y = 'Count') +
  theme(axis.title = element_text(size = 13),
    plot.title = element_text(family = "AvantGarde",
        size = 16, hjust = 0.5, vjust = 0),
    panel.background = element_rect(fill = "gray79",
        colour = "white", linetype = "dotted"),
    plot.background = element_rect(fill = "antiquewhite1"))
general_percent_histogram
```




```
ttl_disb_histogram <- ggplot(data=subset(campaigns, !is.na(ttl_disb)), aes(x=log10(ttl_disb))) +
geom_histogram(bins = 150) +
  labs(
    title = 'Total Campaign Spending', x = 'Spending in $Millions', y = 'Count') +
  scale_x_continuous(breaks=seq(0, 10, 1),labels = 10^(seq(0,10,1)-6)) +
  theme(axis.title = element_text(size = 13),
    plot.title = element_text(family = "AvantGarde",
        size = 16, hjust = 0.5, vjust = 0),
    panel.background = element_rect(fill = "gray79",
        colour = "white", linetype = "dotted"),
    plot.background = element_rect(fill = "antiquewhite1"))

general_percent_histogram / ttl_disb_histogram
```



>

Looking at the voting distribution, we can see that most of the candidates received less than 15% of the total shares of the votes (far left side of the graph). Few of the candidates received the majority of the votes(far right side of the graphs). While the remainder of the candidnates received about 50% +/-(5-10%).

Looking at the total amount of money spent by campaing, we can observe that the majority of the candidates spent close $0.1M to $3M. Very few campaigns spent less than or greater than aforemention amount.

2. Exploring the relationship between spending and votes.

2. (3 points) Create a new dataframe by joining `results_house` and `campaigns` using the `inner_join` function from `dplyr`. (We use the format `package::function` – so `dplyr::inner_join`.)

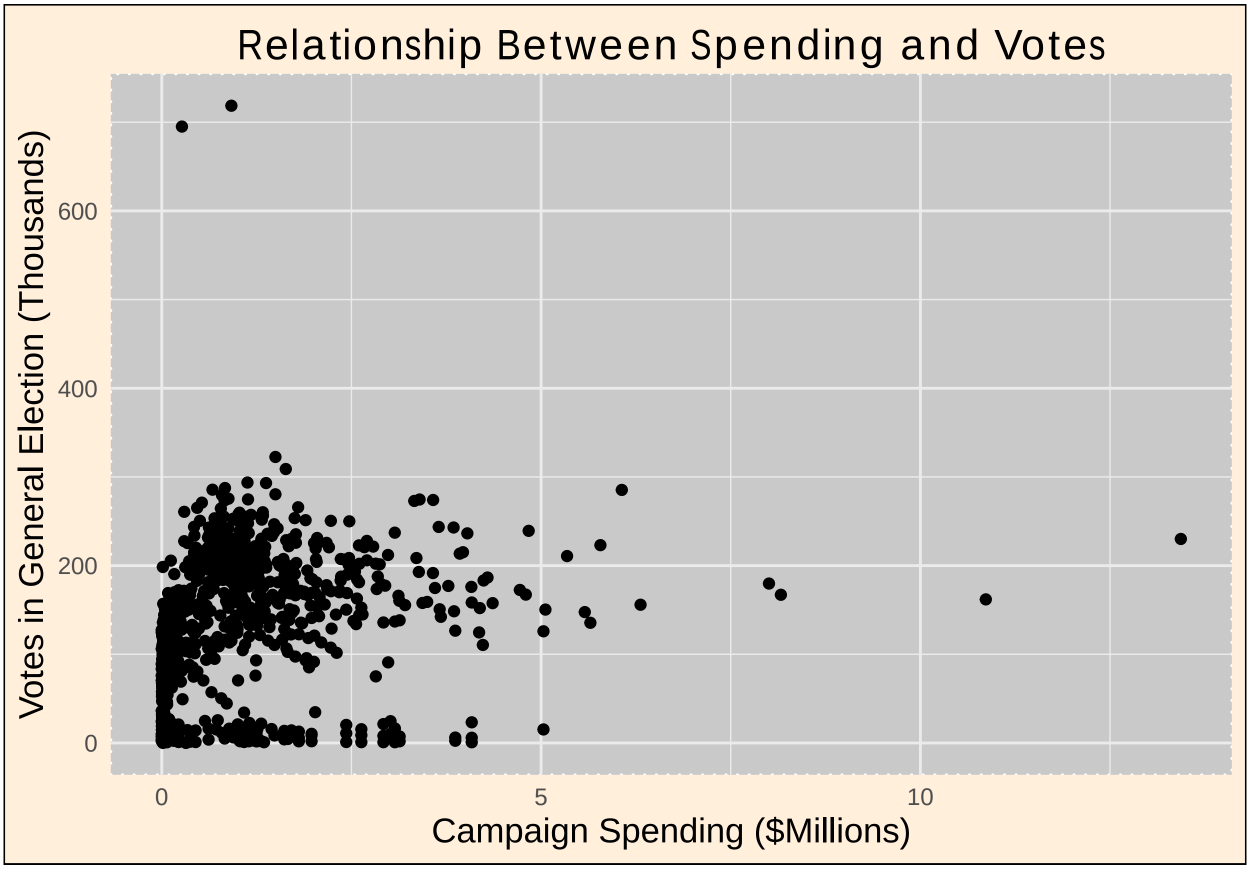
```
results_house_and_campaign <- inner_join(results_house, campaigns, by = "cand_id")
```

3. (3 points) Produce a scatter plot of `general_votes` on the y-axis and `ttl_disb` on the x-axis. What do you observe about the shape of the joint distribution?

```
general_votes_vs_ttl_disb_plot <- results_house_and_campaign %>%
  ggplot() +
  aes(x = ttl_disb/1000000, y = general_votes/1000) +
  geom_point() +
  labs(
    title = 'Relationship Between Spending and Votes',
    x = 'Campaign Spending ($Millions)',
```

```
    y = 'Votes in General Election (Thousands)'
  ) +
  theme(axis.title = element_text(size = 13),
    plot.title = element_text(family = "AvantGarde",
        size = 16, hjust = 0.5, vjust = 0),
    panel.background = element_rect(fill = "gray79",
        colour = "white", linetype = "dotted"),
    plot.background = element_rect(fill = "antiquewhite1"))

general_votes_vs_ttl_disb_plot
```



> Looking at the graph, we can see that there are a lot of campaigns who garnered relatively small number of votes despite spending $millions (below 200,000 votes). A few campaigns, spent over $5 millions but not cracking over 200,000 votes (right side of the graph). Interestingly, we can also observe a very small number of campaigns garnered relatively very larger number of votes (over 600,000 votes) despite spending less than $1 million (top left of the graph).

4. (3 points) Create a new variable to indicate whether each individual is a “Democrat”, “Republican” or “Other Party”.

- Here’s an example of how you might use `mutate` and `case_when` together to create a variable.

```
starwars %>%
  select(name:mass, gender, species) %>%
  mutate(
```

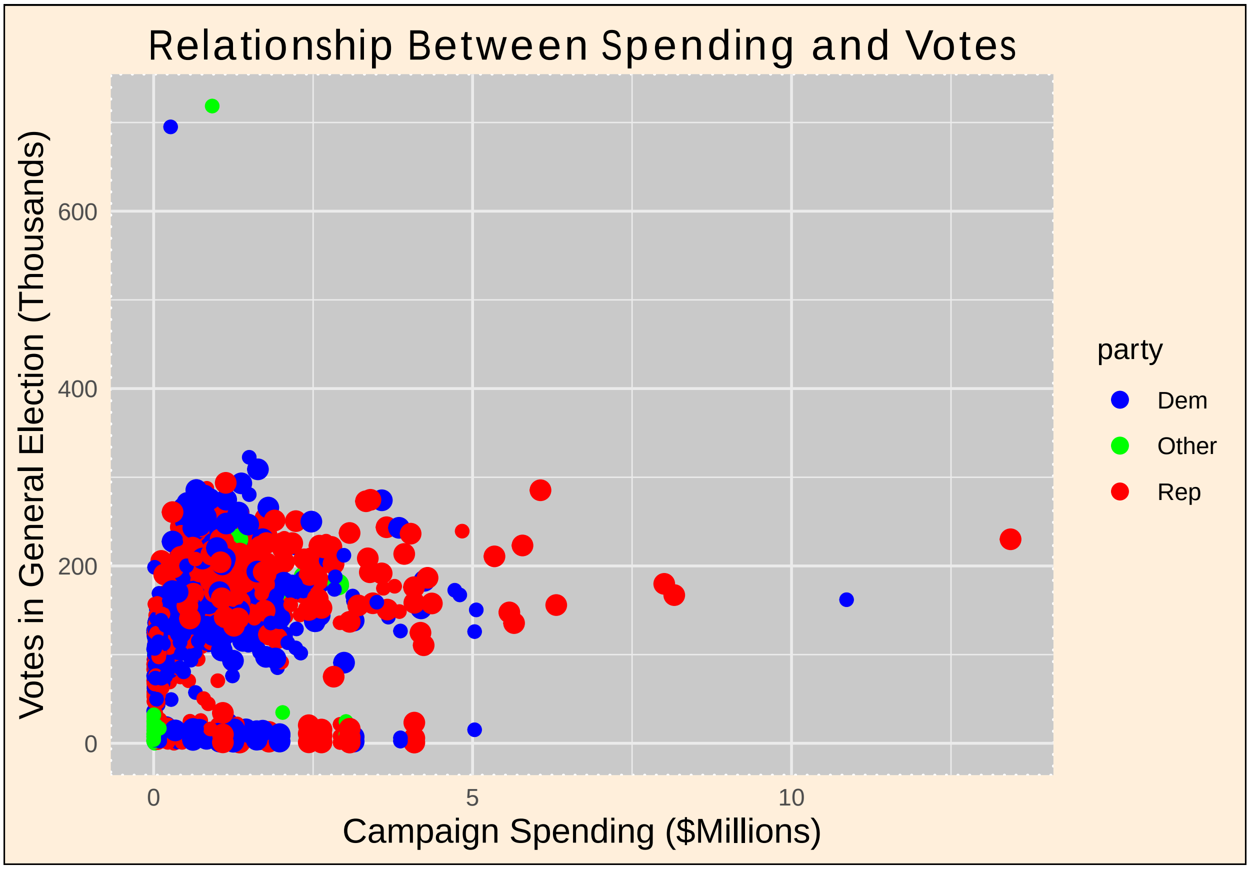
```
  type = case_when(
    height > 200 | mass > 200 ~ "large",
    species == "Droid"        ~ "robot",
    TRUE                      ~ "other"
    )
  )
```

Once you've produced the new variable, plot your scatter plot again, but this time adding an argument into the `aes()` function that colors the points by party membership. What do you observe about the distribution of all three variables?

```
democrat_republican <- results_house_and_campaign %>%
  mutate(party =
           case_when(
             cand_pty_affiliation == "REP" ~ "Rep",
             cand_pty_affiliation == "DEM" ~ "Dem",
             TRUE ~ "Other"
           ))


democrat_republican_plot <- democrat_republican %>%
  ggplot() +
  aes(x = ttl_disb/1000000, y = general_votes/1000) +
  geom_point(aes(col=party, stroke=as.integer(incumbent)), size=2.5) +
  scale_color_manual(values=c("blue", "green", "red")) +
  labs(
    title = 'Relationship Between Spending and Votes',
    x = 'Campaign Spending ($Millions)',
    y = 'Votes in General Election (Thousands)'
  ) +
  theme(axis.title = element_text(size = 13),
    plot.title = element_text(family = "AvantGarde",
        size = 16, hjust = 0.5, vjust = 0),
    panel.background = element_rect(fill = "gray79",
        colour = "white", linetype = "dotted"),
    plot.background = element_rect(fill = "antiquewhite1"))

democrat_republican_plot
```



Looking at the graph, we can see that generally speaking, democrats and republicans got similar number of votes vs campaign spending. A couple of points to notes, it does look like some republicans spent more in campaigning than democrats while attanings about the same number of votes. Interesting point to note that partys other than democrat or republican spent relatively small amount of money while garnering very few votes. Furthermore, there seems to be a democratic campaign and an other party campaign that got the most votes while spending less than $1 million.

Produce a Descriptive Model

5. (5 Points) Given your observations, produce a linear model that you think does a good job at describing the relationship between candidate spending and votes they receive. You should decide what transformation to apply to spending (if any), what transformation to apply to votes (if any) and also how to include the party affiliation.

   For model_1, I will use just the total spending variable and see what results the model will produce.

$$Votes = \beta_0 + \beta_1 \cdot \log(\text{Total disbursements})$$

β_0 will be the votes that a candidate gets with no additional variables. β_1 will be the incremental votes resulting by increasing spending by 1%

```r
results_house_and_campaign <- inner_join(results_house, campaigns, by = "cand_id")

#dropping na values form general_votes column and ttl_disb column
cleaned_data1 <- results_house_and_campaign %>%
  drop_na(general_votes, ttl_disb)

model_1 <- lm(general_votes ~ log(ttl_disb+1), data = cleaned_data1)

summary(model_1)
```

```
##
## Call:
## lm(formula = general_votes ~ log(ttl_disb + 1), data = cleaned_data1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -170802  -34074    7627   45061  568397
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)          -47021      14436  -3.257  0.00117 **
## log(ttl_disb + 1)     14364       1110  12.936  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 73730 on 878 degrees of freedom
## Multiple R-squared:  0.1601, Adjusted R-squared:  0.1591
## F-statistic: 167.3 on 1 and 878 DF,  p-value: < 2.2e-16
```

see question 7 for discussion of model_1

$$Votes = \beta_0 + \beta_1 \cdot Incumbent + \beta_2 \cdot \log(\text{Total disbursements}) + \beta_3 \cdot party$$

β_0 will be the votes that a candidate gets with no additional variables. β_1 votes resulting by increasing spending by 1% β_2 votes that an incumbent gets. β_3 votes resulting for being associated with a party. β_4 votes resulting from the election being in a different States.

```r
#creating a new column by with Dem, Rep, other Designations
#create a new column by designating if the candidate is incumbent or challenger
cleaned_data2 <- cleaned_data1 %>%
  mutate(party_2 =
           case_when(
             cand_pty_affiliation=="REP" ~ "REP",
             cand_pty_affiliation=="DEM" ~ "DEM",
             TRUE                       ~ "Other")
         ) %>%
  mutate(INCU = ifelse(incumbent=="TRUE","Incumbent","Challenger"))

#categorizing incumbent
cleaned_data2$INCU<- as.factor(cleaned_data2$INCU)

model_2 <- lm(general_votes ~ log(ttl_disb+1) + incumbent + party_2 + state, data = cleaned_data2)
summary(model_2)
```

```
##
## Call:
## lm(formula = general_votes ~ log(ttl_disb + 1) + incumbent +
##     party_2 + state, data = cleaned_data2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -401197  -25814   -2496   24468  243393
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)        -16319.0    36321.9  -0.449   0.6533
## log(ttl_disb + 1)    8972.0      959.2   9.353   <2e-16 ***
## incumbentTRUE       49037.0     4052.5  12.101   <2e-16 ***
## party_2Other       -73189.5     8080.7  -9.057   <2e-16 ***
## party_2REP          -1541.6     3439.7  -0.448   0.6541
## stateAL             40850.0    37120.6   1.100   0.2715
## stateAR             36968.8    40419.4   0.915   0.3607
## stateAS            -87797.1    44156.1  -1.988   0.0471 *
## stateAZ             23945.4    36328.6   0.659   0.5100
## stateCA              9687.5    34467.7   0.281   0.7787
## stateCO             58081.6    36475.4   1.592   0.1117
## stateCT            -31338.1    36334.0  -0.863   0.3887
## stateDC            115346.4    59162.3   1.950   0.0516 .
## stateDE            101152.5    48268.1   2.096   0.0364 *
## stateFL             47165.5    34721.2   1.358   0.1747
## stateGA             55076.2    36028.8   1.529   0.1267
## stateGU            -97731.4    48276.1  -2.024   0.0433 *
## stateHI             27963.8    44068.8   0.635   0.5259
## stateIA             54883.3    38134.7   1.439   0.1505
## stateID             51056.2    41846.1   1.220   0.2228
## stateIL             44334.0    35215.8   1.259   0.2084
## stateIN             33749.6    36095.3   0.935   0.3501
## stateKS             -5803.4    37719.4  -0.154   0.8778
## stateKY             60548.2    37105.6   1.632   0.1031
## stateLA            -29459.1    35888.0  -0.821   0.4120
## stateMA             70938.5    36235.2   1.958   0.0506 .
## stateMD             52115.7    36470.5   1.429   0.1534
## stateME             56936.0    41776.0   1.363   0.1733
## stateMI             43300.5    35307.7   1.226   0.2204
## stateMN             65808.3    36368.2   1.810   0.0707 .
## stateMO             61220.3    36677.1   1.669   0.0955 .
## stateMP            -47091.1    59656.0  -0.789   0.4301
## stateMS             44544.3    39528.7   1.127   0.2601
## stateMT            101707.9    48250.3   2.108   0.0353 *
## stateNC             62284.1    35489.1   1.755   0.0796 .
## stateND             35939.2    44127.0   0.814   0.4156
## stateNE             30205.6    41796.1   0.723   0.4701
## stateNH             -8472.9    40391.8  -0.210   0.8339
## stateNJ             26058.4    35661.0   0.731   0.4652
## stateNM              3585.3    40367.3   0.089   0.9292
## stateNV              4418.7    37720.0   0.117   0.9068
## stateNY            -75693.0    34394.1  -2.201   0.0280 *
## stateOH             48200.4    35270.0   1.367   0.1721
```

```
## stateOK             34703.3    40400.0   0.859   0.3906
## stateOR             77168.4    38695.6   1.994   0.0465 *
## statePA             61812.1    35269.9   1.753   0.0801 .
## statePR            441522.8    44323.2   9.961   <2e-16 ***
## stateRI            -11608.2    41829.6  -0.278   0.7815
## stateSC             -3534.1    35994.6  -0.098   0.9218
## stateSD             52654.0    48236.3   1.092   0.2753
## stateTN             37137.7    36560.4   1.016   0.3100
## stateTX             12744.2    34773.6   0.366   0.7141
## stateUT              5139.2    38137.6   0.135   0.8928
## stateVA             57691.1    35704.0   1.616   0.1065
## stateVI           -132421.1    59173.1  -2.238   0.0255 *
## stateVT            109983.7    59148.5   1.859   0.0633 .
## stateWA             38716.5    36196.3   1.070   0.2851
## stateWI             59074.0    36099.2   1.636   0.1021
## stateWV             -6488.8    38704.8  -0.168   0.8669
## stateWY             23016.2    44132.0   0.522   0.6021
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 48240 on 820 degrees of freedom
## Multiple R-squared:  0.6643, Adjusted R-squared:  0.6401
## F-statistic:  27.5 on 59 and 820 DF,  p-value: < 2.2e-16
```

see question 7 for discussion of model_2

6. (3 points) Evaluate the Large-Sample Linear Model Assumptions

   Large-Sample Linear Models have two assumptions; Unique BLP exits and I.I.D Since we have a large enough sample, it is very likely that a unique BLP exists, satisfying our first assumption. Although for our state coefficient, one can argue that neighboring states affect each other, rendering I.I.D. as obsolete for us. However, we will assume that each state has no effect on the other and we would assume that they are independent of each other.

7. (3 points) Interpret the model coefficients you estimate.

   After running our linear regression, we can see that the p-value is significant and we can reject the null hypothesis that the total spending has no effect on number of votes. In other words, total spending(ttl_disb) has a significant effect on the number of votes. Inreasing spending by 1% increases the number of votes on average by 14364 votes.

   let's see if we can make our model fit better by introducing three more coefficients; incumbuncy, party, and state.

   Looks like model_2 does a better job (note: introducing more coefficients may make our model even better, I decided to just use the above 4 coefficient for this HW). Looking at the regression, I will list some of the interesting things that exists in the above table.

- R-square increased in our second model as expected.
- Total spending coefficient, incumbency are all significant. Therefore, it is safe to say each one has an effect on the number of votes. As for the states, some of them are significant and some are not, but jointly they do improve the adjusted R-square to ~ 65%.

- looks like being an incumbent will garner an additional ~49,000 votes
- Being independent can cost ~73,000 votes and and is significant
- Being a democrat or republican is not significant and fail to affect the outcome.
- Tasks to keep in mind as you're writing about your model:
  - At the time that you're writing and interpreting your regression coefficients you'll be *deep* in the analysis. Nobody will know more about the data than you do, at that point. *So, although it will feel tedious, be descriptive and thorough in describing your observations.*
  - It can be hard to strike the balance between: on the one hand, writing enough of the technical underpinnings to know that your model meets the assumptions that it must; and, on the other hand, writing little enough about the model assumptions that the implications of the model can still be clear. We're starting this practice now, so that by the end of Lab 2 you will have had several chances to strike this balance.