R Camp - Day 1 Homework

Emily Elia

8/18/2019

```
setwd("/Users/etelia/Documents")
load(file="RDAY1_Data.RData")
```

Day 1: Data Exploration

Dataset Used

My observational data sets comes from a 2015 paper by Pablo Fernandez-Vazquez, Pablo Barbera, and Gonzalo Rivero titled, "Rooting Out Corruption or Rooting for Corruption? The Heterogeneous Electoral Consequences of Scandals. This paper investigates the impact of corruption scandals on vote share during the 2011 local elections in Spain. The Spanish housing boom generated numerous corruption scandals within local Spanish government. The authors look at the vote share of mayors from 2007 to 2011 (in between which the housing boom scandals occured) in order to determine how different types of corruption and their media coverage impact electoral gains and losses.

Using this dataset, I examine whether or not the presence of a majority government impacts the difference in vote share from 2007 to 2011 when corruption occurs. Will corrupt mayors in a majority government suffer greater vote loss than corrupt mayors in a non-majority government? According to Schwindt-Bayer & Tavits 2015, citizens are more likely to electorally punish corrupt politicians when there is a majority government because clarity of responsibility is high. A majority government makes it clear which party and which politicians are to blame for acts of corruption, so citizens are more likely to know which politicians and which party to vote out of office. Therefore, I anticipate corrupt mayors possessing a majority government at the time of elections in 2011 to suffer a greater loss of votes than corrupt mayors in a non-majority government and than noncorrupt mayors in either a majority or non-majority government.

Problem 1

Summarizing Data

Majority Govt + Corruption: abs_maj

A dichotomous variable where 1 = mayor that committed corruption and has majority government, and 0 = all other mayors.

abs_maj	
Mean	0.8
Median	1
Min	0
Max	1

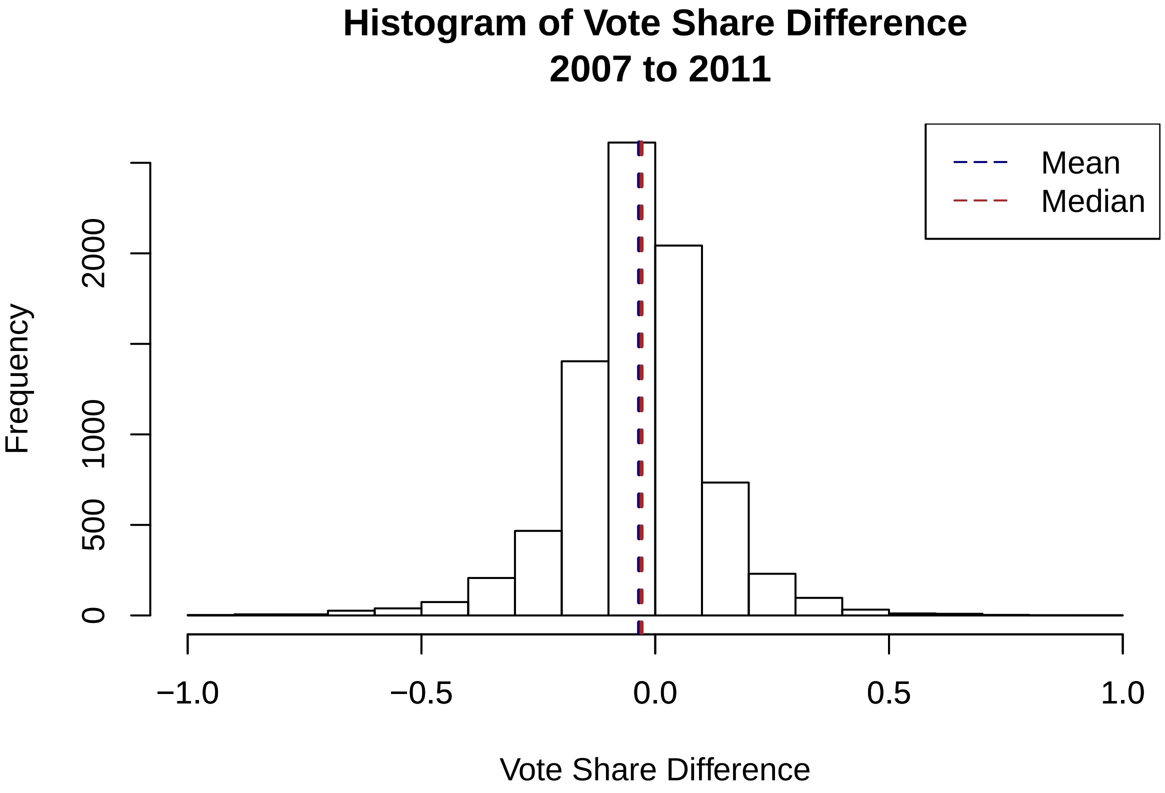
Corruption: corruption

A dichotomous variable where 1 = known corruption occurred (covered by media sources) and 0 = no known corruption ocurred

Vote Share Difference (percentage):

sharediff	
Mean	-0.035
Median	-0.029
Min	-0.927
Max	0.940

```r
hist(sharediff, xlab="Vote Share Difference", ylab="Frequency", main="Histogram of Vote Share Differen
axis(side=1, at=seq(-1.0, 1.0, by=.5))
abline(v=mean(sharediff, na.rm=TRUE),   col="navy", lwd="2", lty=2)
abline(v=median(sharediff, na.rm=TRUE), col="firebrick", lwd="2", lty=2)
legend("topright", legend=c("Mean", "Median"),
     col=c("navy", "firebrick"), lty=c(5,5))
```



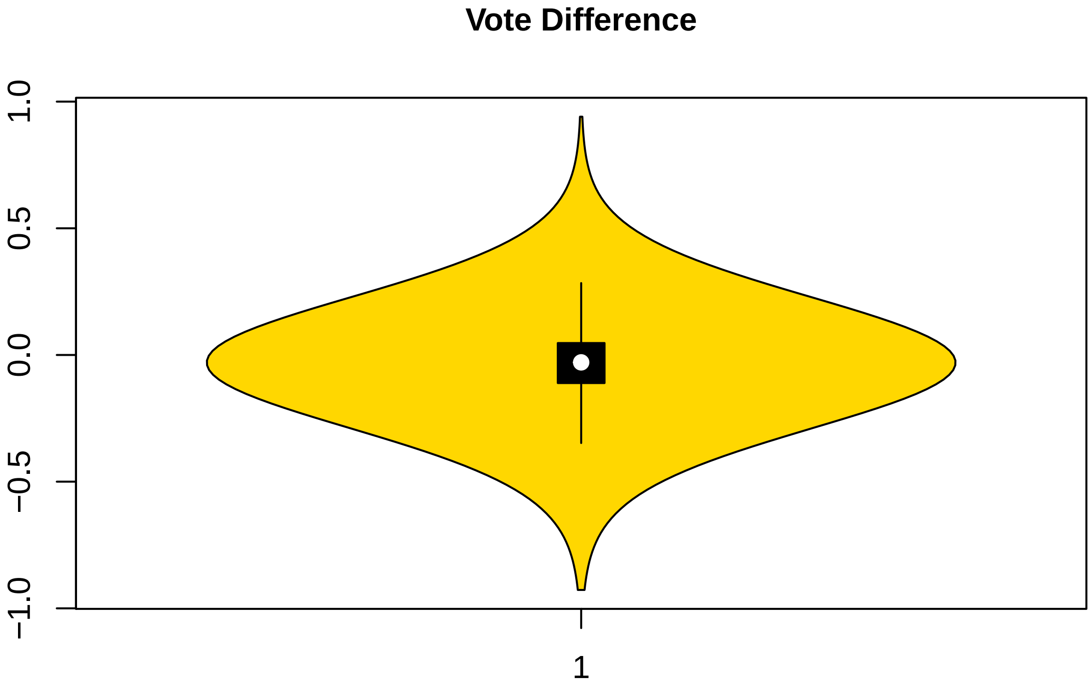
```r
library(vioplot)
```

```
## Warning: package 'vioplot' was built under R version 3.5.2

## Loading required package: sm

## Package 'sm', version 2.2-5.4: type help(sm) for summary information

## Loading required package: zoo

##
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':
##
```

```
##     as.Date, as.Date.numeric
```

```
vioplot(sharediff, main="Vote Difference", col="gold")
```

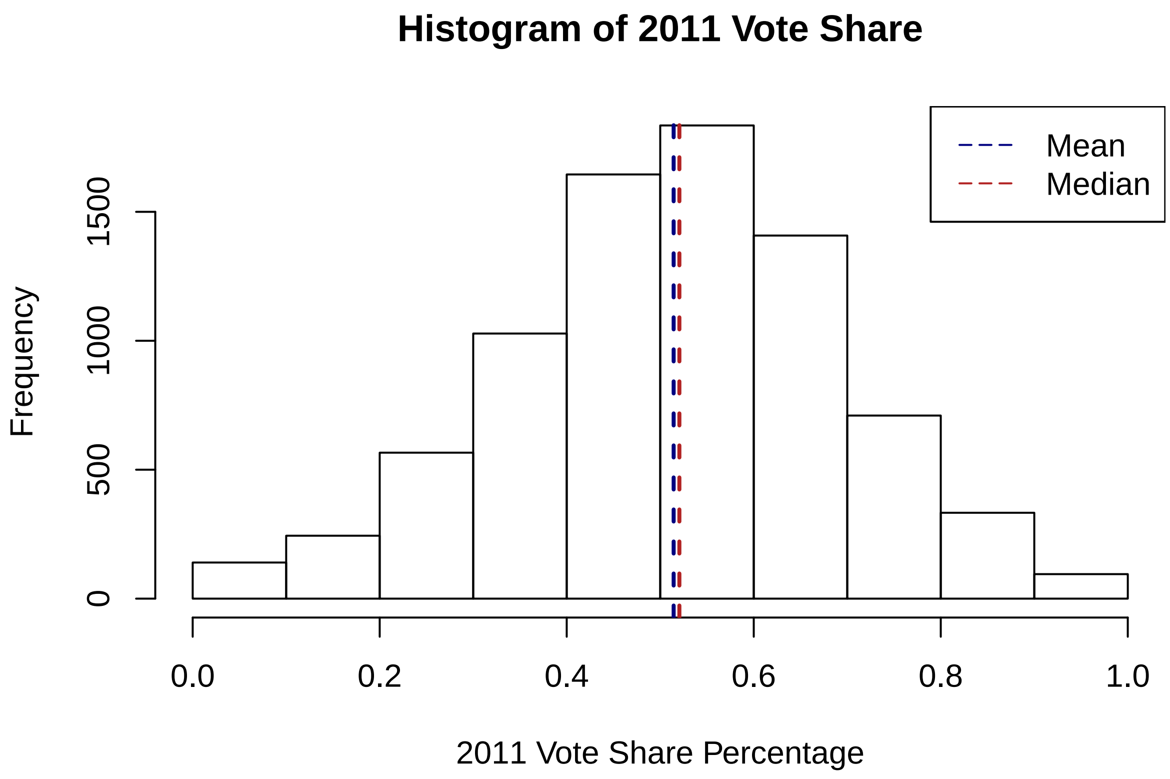
```
## [1] -0.9273615  0.9404410
```



2011 Vote Share (percentage):

percent_curr	
Mean	0.51
Median	0.52
Min	0.0027
Max	1.00

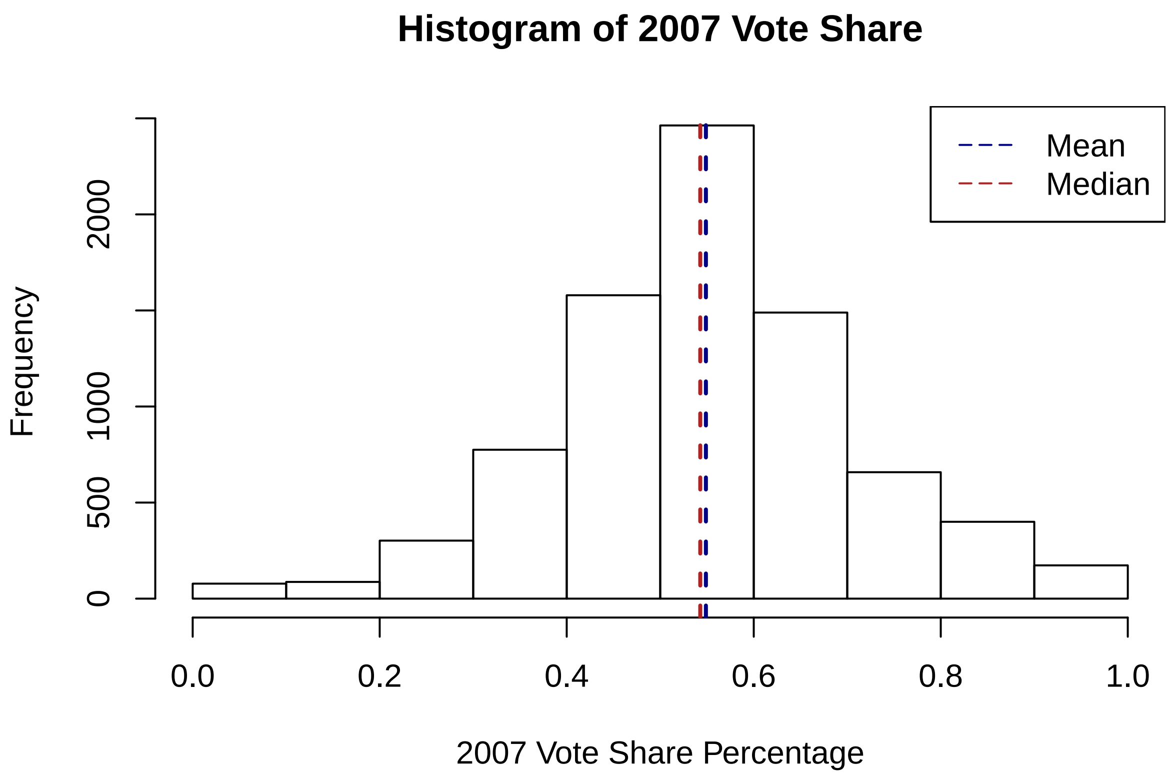
```
hist(maindat$percent_curr, xlab="2011 Vote Share Percentage", main="Histogram of 2011 Vote Share")
  abline(v=mean(maindat$percent_curr, na.rm=TRUE),   col="navy", lwd="2", lty=2)
  abline(v=median(maindat$percent_curr, na.rm=TRUE), col="firebrick", lwd="2", lty=2)
  legend("topright", legend=c("Mean", "Median"),
       col=c("navy", "firebrick"), lty=c(5,5))
```



2007 Vote Share (percentage):

percent_prev	
Mean	0.55
Median	0.54
Min	0.0013
Max	1.00

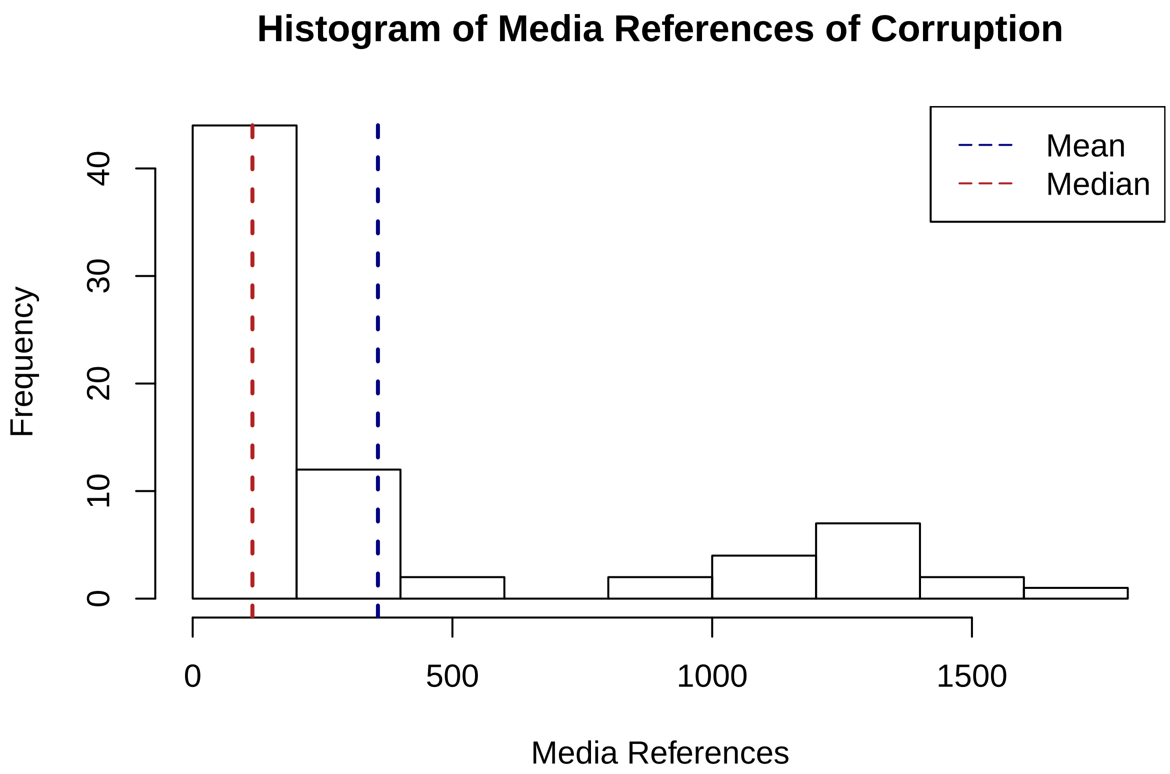
```
hist(maindat$percent_prev, xlab="2007 Vote Share Percentage", main="Histogram of 2007 Vote Share")
  abline(v=mean(maindat$percent_prev, na.rm=TRUE),   col="navy", lwd="2", lty=2)
  abline(v=median(maindat$percent_prev, na.rm=TRUE), col="firebrick", lwd="2", lty=2)
  legend("topright", legend=c("Mean", "Median"),
       col=c("navy", "firebrick"), lty=c(5,5))
```



Media References of Corruption:

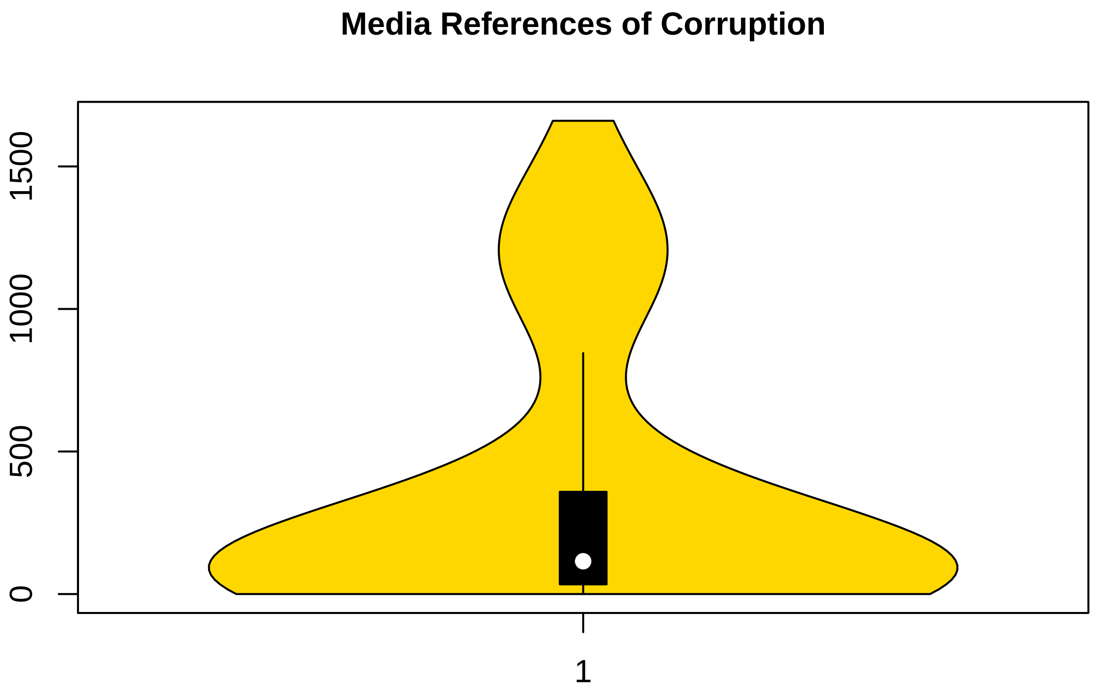
references	
Mean	356.608
Median	115.00
Min	0
Max	1660

```
hist(maindat$references, xlab="Media References", main="Histogram of Media References of Corruption")
  abline(v=mean(maindat$references, na.rm=TRUE),   col="navy", lwd="2", lty=2)
  abline(v=median(maindat$references, na.rm=TRUE), col="firebrick", lwd="2", lty=2)
  legend("topright", legend=c("Mean", "Median"),
     col=c("navy", "firebrick"), lty=c(5,5))
```



```
vioplot(maindat$references, main="Media References of Corruption", col="gold")
```

```
## [1]    0 1660
```



"Welfare-Enhancing" Corruption:

Corruption that gives some benefit to the voter and thus is less harshly judged

w_enhance	
Mean	0.0036
Median	0.00
Min	0.00
Max	1.00

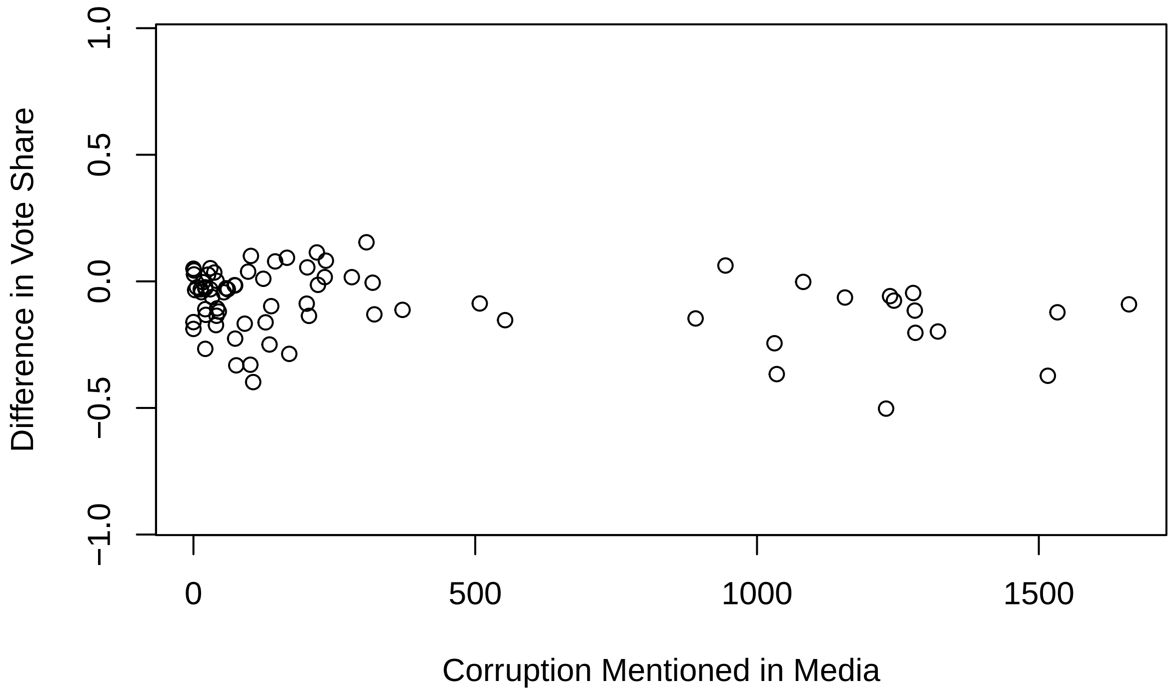
“Welfare-Decreasing” Corruption:

Corruption that gives no benefit to the voter and/or harms the voter and thus is more harshly judged

w_decrease	
Mean	0.0057
Median	0.00
Min	0.00
Max	1.00

Relationship Between Vote Share Difference and Media References of Corruption:

```
plot(maindat$references, sharediff, xlab="Corruption Mentioned in Media", ylab="Difference in Vote Sha
```



Discussion of Variables

The variables behave the way I expect them to based on theoretical assumptions. The 2007 vote share, 2011 vote share, and the difference in vote share all have means and medians that I would expect, and their distributions are not skewed. The biggest challenge with the dataset is that it is impossible to know how many mayors committed corruption but were able to hide it. However, this lack of information is insignificant when focusing vote share losses and gaines because if corruption occured without detection, then voters would not be aware of it and thus it would have no impact on their voting behavior.

The variable that departs from expectations most drastically is the references variable, which refers to the number of times in which media covered a specific mayor's corruption scandal. The bulk of the data is between 0 and 500 references, but then there is also a sizeable amount of data points between ~1000 and ~1700 references. Clearly, a select amount of corruption cases got heightened media attention. The amount of media attention can influence voter behavior, so I control for references in my models.

Problem 2

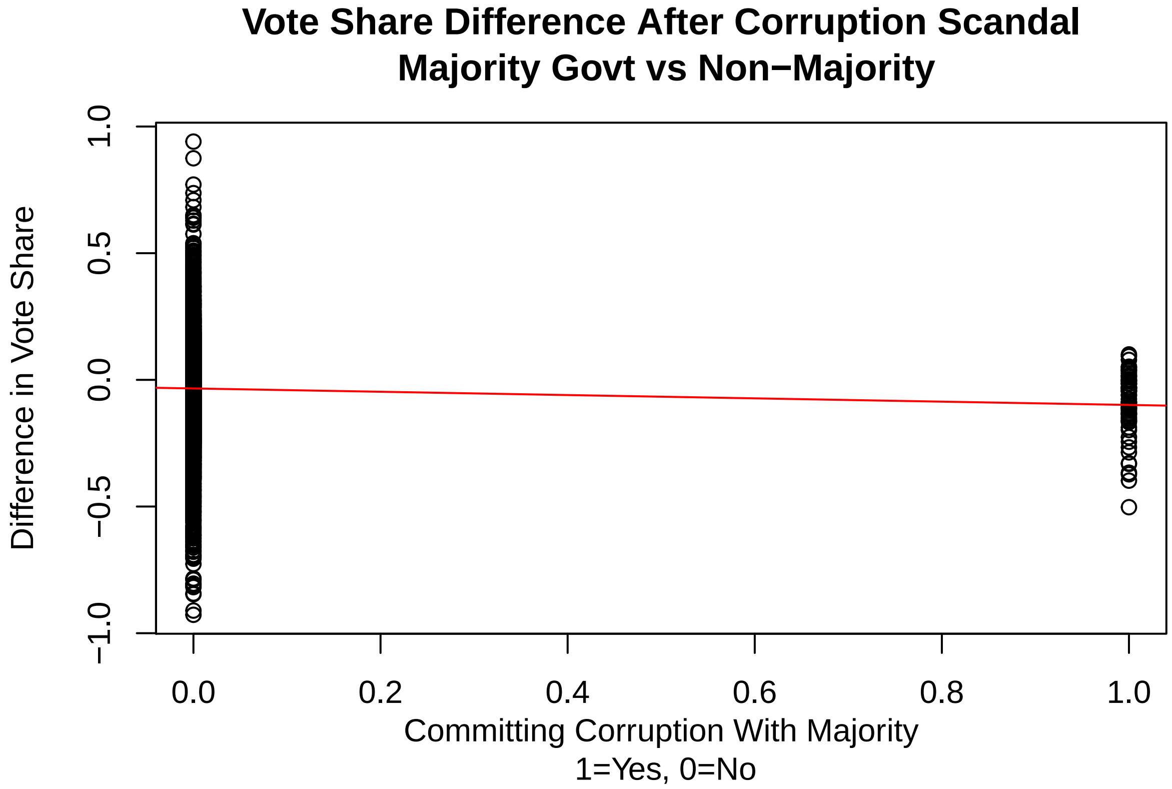
Exploring Models

Because my dependent variable, vote share difference, can take on both positive and negative values, I use a normal linear regression model to explore the relationship between majority government and vote share difference. However, the dependent variable is theoretically and numerically bounded between -1 and 1, since the variable is a percentage of vote share. A person cannot lose more than 100% of the vote share nor can they win more than 100% of the vote share.

I debated taking the log of vote share to have only positive values, but doing so would not make theoretical sense because it would eliminate the advantage of clearly viewing changes in vote share in both the negative and positive direction. By allowing negative values, I can easily identify who lost votes and avoid any potential confusion with observing which mayors simply had the largest change in vote share without any distinction of positive vs. negative value, or gains vs. losses.

Scatter Plot of sharediff over maj_corr with regression line

```
plot(maj_corr, sharediff, xlab="Committing Corruption With Majority\n 1=Yes, 0=No",
     ylab="Difference in Vote Share")
    title("Vote Share Difference After Corruption Scandal\n Majority Govt vs Non-Majority")
abline(m_ols, col="red")
```



Normal-Linear Model

Regression Tables

Normal-Linear Model with no controls Effect of having Majority Govt over Vote Share Difference

```r
m_ols <- lm(sharediff ~ maj_corr, data=maindat)

summary(m_ols)
```

```
##
## Call:
## lm(formula = sharediff ~ maj_corr, data = maindat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.89334 -0.07625  0.00529  0.08225  0.97446
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.03402    0.00172 -19.781  < 2e-16 ***
## maj_corr    -0.06506    0.01986  -3.276  0.00106 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1533 on 8002 degrees of freedom
##   (6 observations deleted due to missingness)
## Multiple R-squared:  0.001339,   Adjusted R-squared:  0.001214
## F-statistic: 10.73 on 1 and 8002 DF,  p-value: 0.001059
```

Normal-Linear Model with controls Effect of having Majority Govt over Vote Share Difference Controlling for kind of corruption, population change from 2007 to 2011, and amount of media references to the corruption scandal

```r
m_ols_c <- lm(sharediff ~ maj_corr + maindat$references + popdiff + maindat$w_enhance + maindat$w_decrea

summary(m_ols_c)
```

```
##
## Call:
## lm(formula = sharediff ~ maj_corr + maindat$references + popdiff +
##     maindat$w_enhance + maindat$w_decrease, data = maindat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.35559 -0.07324  0.02065  0.09475  0.18988
##
## Coefficients: (1 not defined because of singularities)
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        -1.932e-02  3.623e-02  -0.533   0.5957
## maj_corr           -6.408e-02  3.703e-02  -1.730   0.0882 .
## maindat$references -4.448e-05  3.667e-05  -1.213   0.2294
## popdiff            -5.575e-06  4.868e-06  -1.145   0.2562
## maindat$w_enhance   2.982e-02  3.153e-02   0.946   0.3476
## maindat$w_decrease         NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1258 on 67 degrees of freedom
##   (7938 observations deleted due to missingness)
## Multiple R-squared:  0.152,  Adjusted R-squared:  0.1013
## F-statistic: 3.002 on 4 and 67 DF,  p-value: 0.02434
```

Discussion

To what extent is linear regression helpful in discovering the theoretical functional form?

In order to discover the best theoretical functional form for a model, I believe the best route is to first determine what theoretical limits your model ought to have based on the data generating process and theory before running any kind of model. For example, determining the theoretical range of values - Can my values be negative? Are my values between 0 and 1? - will have a major impact on model choice. Running a linear regression as part of the process to discover the theoretical form can be misleading, because your model will fit the data to a linear relationship even if a linear relationship doesn't truly exist. If you run a linear model before giving thought to whether or not this relationship may actually be linear, you may inadvertently influence your model-building process. I think running a linear regression should be done only after the best theoretical model is determined in an abstract sense based on data and theory before one even takes up running a statistical model in code.