Simran Mallik

QTM L Section 11

Analysis of the 2016 Presidential Election

Introduction

600 word limit

The 2016 presidential election was a historical election, and may have potentially had the effect of changing voting patterns observed in prior elections. This project examines US county voting patterns and can be used to highlight problems in voting and society. The association of a county's population in relation to the national county average number of votes is examined; the number of people in a county may impact the number of votes (positive/negative association). I hypothesize larger counties cast more votes. The average proportion of the total number of votes (# of votes in county/total number of popular election votes) of majority white counties is compared to that of majority non-white counties, which may be associated due to different voting behaviors. I hypothesize majority white counties will vote more than majority nonwhite counties. Lastly, the correlation between proportion of total number of votes per county and percentage of females per county is examined. Women have different interests and voting behaviors than men, and I hypothesize higher vote counts in counties with greater percentages of females.

Methods

The dataset contains US county voting statistics. I used the number of votes per county variable, removed nonsensical values (like 99999999), and recoded it into a proportion of votes per county variable (total votes per county/139 million, total votes in popular election (**Gould, Skye, 2016**)). I used the percentage of females per county variable and removed nonsensical values (counties 100% female and counties less than 25% female). I recoded the percentage of nonwhite inhabitants variable into two categories: "Majority nonwhite" and "Majority white". Another variable categorizes counties into population groups by assigning numbers to counties with specific populations. I consolidated two groups in the new variable, and removed nonsensical values. I found the mean of the proportion of total votes in each county

and created a new variable categorizing counties into “More votes” and “Less votes” than national county average.

Results

(1) There’s an association between county population and number of votes casted compared to national county average number of votes. Since the test statistic p value is small, we conclude there’s an association.

(2) There’s a statistical difference between the average of the proportion of total number of votes for majority white counties and that of majority nonwhite counties. Since the test statistic p value is significantly small, there’s an association. The mean proportion of total votes in counties majority white (0.0007608111) is larger than that of counties majority nonwhite (0.0002590252).

(3) There’s a statistically significant correlation between the county proportion of the total number of votes and the percentage of females. Since the test statistic p value is less than alpha, we conclude there’s a correlation, which is 0.129783 (weak positive).

Discussion

(1) The association is practically significant; further studies could investigate association direction, and significance behind a positive or negative relationship. However, it’s possible we falsely rejected the null (Type 1 error).

(2) The mean in the majority white group being higher than mean in the majority nonwhite group is practically significant. This could result from a higher number of majority white counties, or, for example, there are the same number of majority white and majority nonwhite counties and differences stem from factors like discrimination deterring voters. A Type 1 error could’ve occurred.

(3) There exists a correlation between the percent of females and proportion of total votes in each county. However, since there’s a weak positive correlation, 0.1297839, the results aren’t practically significant.

The distribution of percent of females is approximately normal but the distribution of the proportion of total votes is not, which could cause p-value variability. A Type 1 error could've occurred.

Appendix A

[20 points]: Tables of results

Table 1. Descriptive statistics for Elections_2016. N = 3114

	Overall	Total_votesx (11 NA's)	
		More Votes in County than National County Avg	Less Votes in County than National County Avg
	N = 3098 (removed 16 NA's)	N = 592	N = 2511
female_pctn	50.00 +/- 2.26	50.89 +/- 0.89	49.79 +/- 1.43
nonwhite_pcts Majority Nonwhite County Majority White County	 2757(88.99%) 357(11.52%)		
ruralurban_ccn level 1 level 2 level 3 level 4 level 5 level 6 level 7 level 8	 433(13.98%) 374(12.07%) 355(11.46%) 214(6.91%) 89 (2.87%) 593(19.14%) 428(13.82%) 623(20.11%)		

Table 2. Chi Square Results: total_votesx by ruralurban_ccn [Elections_2016]. N = 3098 (16 NA's total in table)

	More Votes in County than National County Avg	More Votes in County than National County Avg	Test Statistic (x^2)	P-val
	N = 592	N = 2511		
ruralurban_ccn			1150.30	p < 2.20e-16
1	254(42.91%)	178(5.75%)		
2	182(30.74%)	191(62.65%)		
3	133(22.47%)	220(71.01%)		
4	15 (2.53%)	199(6.42%)		
5	6 (1.01%)	83 (2.68%)		
6	0 (0.00%)	591(19.08%)		
7	0 (0.00%)	426(13.75%)		
8	0 (0.00%)	620(20.01%)		

Table 3: 2 Sample T test results: total_votesf by nonwhite_pcts. N = 3103 (11 NA's total in table)
total_votesf = Total Votes in County/Total Votes in Popular Election

	total_votesf	Test Stat	P-val
nonwhite_pcts		-4.52	8.43e-06
Majority of County is Nonwhite	2.59e-4 +/- 5.80e-4		
Majority of County is White	7.61e-4 +/- 2.08e-3		

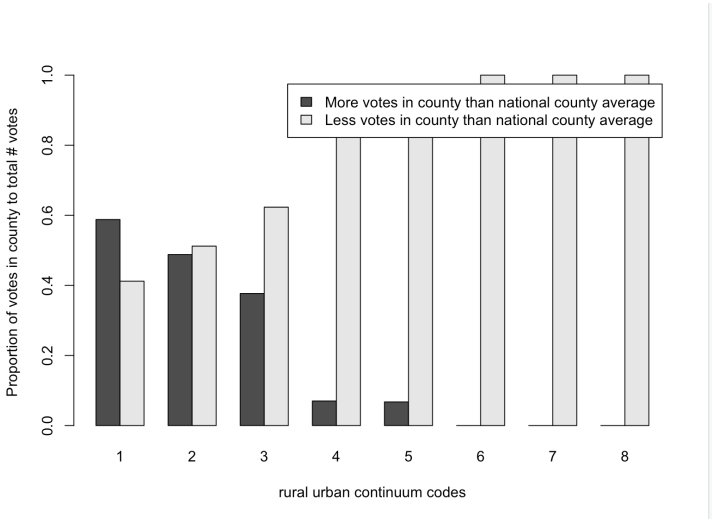
Table 4: Linear Regression Results of total_votesf regressed on female_pctn in Elections_2016.

	total_votesf	95% confidence interval	p-value
	B(se)		
		Lower Upper	
explanatory variable	female_pctn(7.162e-06)	(-2.99e-3, -1.58e-3)	
Intercept/constant	intercept(3.59e-04)	(3.81e-5, 6.61e-5)	
# of Observations	#3091(23 NA's)		
R^2	1.68e-2		
F Statistic	52.92		4.38e-13

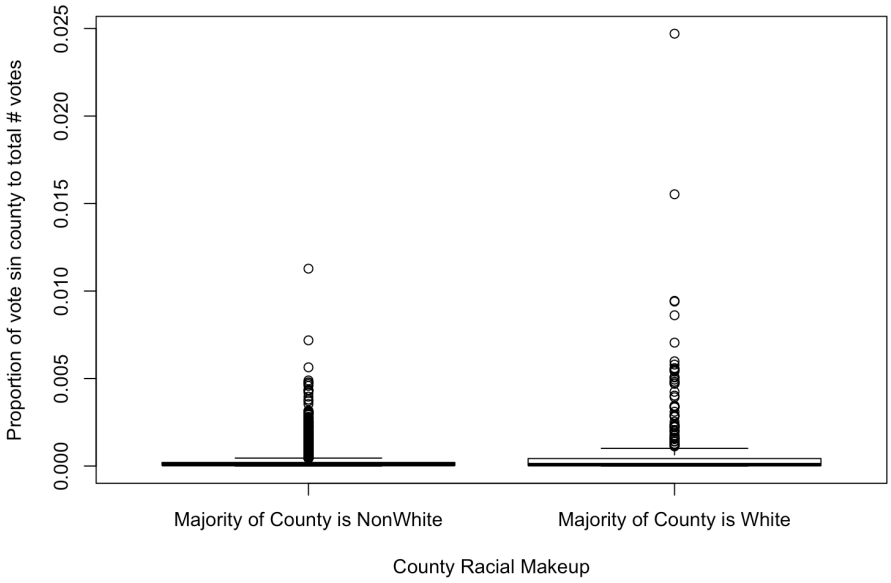
Appendix B:

Bi-variate plots

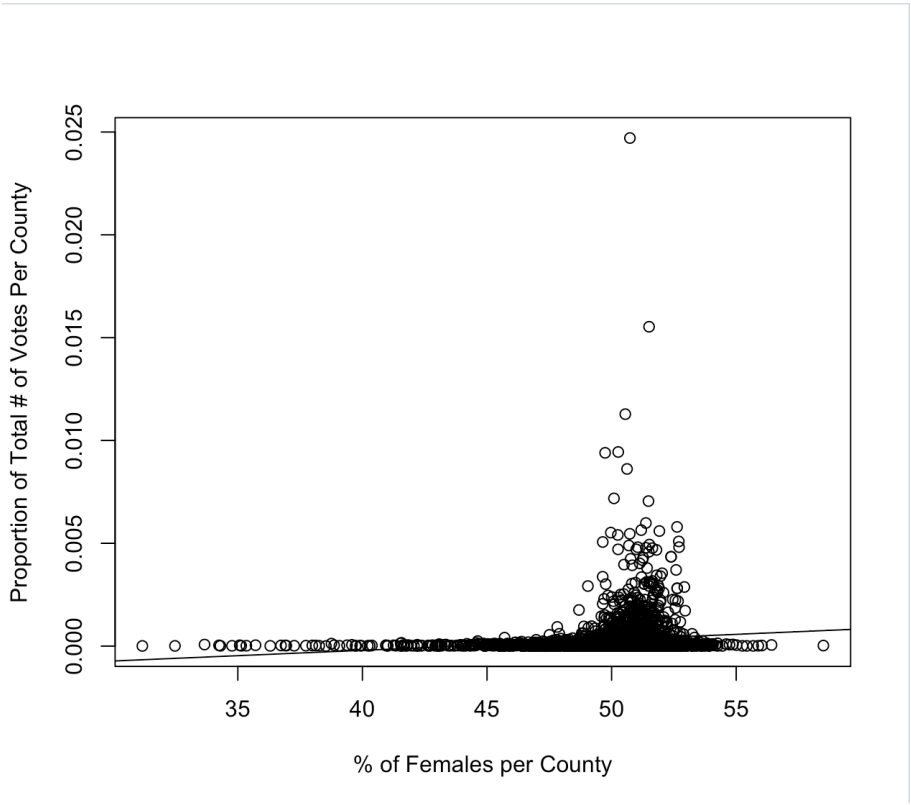
County Population vs Proportion of Votes in County



County Racial Makeup vs Proportion of votes in County



% of Females in County vs Proportion of Votes in County



Appendix C

[6 points]: R code

You should also submit your R code as a separate R file.

```
#1.
#saved dataset on desktop, opened dataset from desktop

summary(Elections_2016)
Elections_2016$total_votes
summary(Elections_2016$total_votes)


#quantitative response variable: total_votesf
#displays proportion of total number of votes per county in a vector
#total_votesf = # of votes in county / total votes casted in popular election
Elections_2016$total_votesn <- Elections_2016$total_votes
Elections_2016$total_votesn[Elections_2016$total_votes == 999999999] <- NA
Elections_2016$total_votesn[Elections_2016$total_votes < 0] <- NA
summary(Elections_2016$total_votesn)
Elections_2016$total_votesf <- Elections_2016$total_votesn/139000000
Elections_2016$total_votesf



#quantitative explanatory variable: female_pctn
#displays percentage of females in population of each county
summary(Elections_2016$female_pct)
Elections_2016$female_pctn <- Elections_2016$female_pct
Elections_2016$female_pctn[Elections_2016$female_pct == 999999999] <- NA
Elections_2016$female_pctn[Elections_2016$female_pct > 100] <- NA
Elections_2016$female_pctn[Elections_2016$female_pct < 25] <- NA
summary(Elections_2016$female_pctn)


#1 categorical (2 levels) explanatory variable: nonwhite_pcts
#displays whether the majority of a county is white or majority of a county is nonwhite
Elections_2016$nonwhite_pcts <- Elections_2016$nonwhite_pct
Elections_2016$nonwhite_pcts <- factor(NA, levels = c("Majority of County is NonWhite", "Majority of
County is White"))
```

```
Elections_2016$nonwhite_pcts[Elections_2016$nonwhite_pct <= 50]  <- "Majority of County is NonWhite"
Elections_2016$nonwhite_pcts[Elections_2016$nonwhite_pct > 50] <- "Majority of County is White"
summary(Elections_2016$nonwhite_pcts)

summary(Elections_2016$nonwhite_pct)

#1 categorical (between 3 and ~8 levels) explanatory variable: ruralurban_ccn
#displays levels associated with the number of ppl in each county
summary(Elections_2016$ruralurban_cc)
Elections_2016$ruralurban_ccs <- Elections_2016$ruralurban_cc
Elections_2016$ruralurban_ccs[Elections_2016$ruralurban_cc == 999999999] <- NA
Elections_2016$ruralurban_ccn <- factor(NA, levels = c("1", "2", "3", "4", "5", "6", "7", "8"))
Elections_2016$ruralurban_ccn[Elections_2016$ruralurban_ccs == 9] <- "8"
Elections_2016$ruralurban_ccn[Elections_2016$ruralurban_ccs == 8] <- "8"
Elections_2016$ruralurban_ccn[Elections_2016$ruralurban_ccs == 7] <- "7"
Elections_2016$ruralurban_ccn[Elections_2016$ruralurban_ccs == 6] <- "6"
Elections_2016$ruralurban_ccn[Elections_2016$ruralurban_ccs == 5] <- "5"
Elections_2016$ruralurban_ccn[Elections_2016$ruralurban_ccs == 4] <- "4"
Elections_2016$ruralurban_ccn[Elections_2016$ruralurban_ccs == 3] <- "3"
Elections_2016$ruralurban_ccn[Elections_2016$ruralurban_ccs == 2] <- "2"
Elections_2016$ruralurban_ccn[Elections_2016$ruralurban_ccs == 1] <- "1"

summary(Elections_2016$ruralurban_ccn)
Elections_2016$ruralurban_ccn
#3.


#dichotomized categorical variable version of response variable: Elections_2016$total_votesx
#displays whether a county has proportion of votes that's greater than the average proportion of votes for all counties in the United States
#of if county has proportion of votes that's less than the average proportion of votes for all counties in the United States

summary(Elections_2016$total_votesf)
Elections_2016$total_votesx <-Elections_2016$total_votesf
Elections_2016$total_votesx <- factor(NA, levels = c("More votes in county than national county average", "Less votes in county than national county average"))
Elections_2016$total_votesx[Elections_2016$total_votesf > 0.000317] <- "More votes in county than national county average"
Elections_2016$total_votesx[Elections_2016$total_votesf <= 0.000317] <- "Less votes in county than national county average"
summary(Elections_2016$total_votesx)
```

#RESEARCH QUESTION

#Question 1: Is the population of a county associated with whether a county has more or less votes than the national county average number of votes?
#plot of dichotomous response variable and one of categorical explanatory variables
plot(Elections_2016$total_votesx)
plot(Elections_2016$ruralurban_ccn)

#Question 2: Does the average proportion of the total number of votes for counties that are majority white differ from counties that are majority nonwhite?
#plot of quantitative response variable and a categorical explanatory variable
boxplot(Elections_2016$total_votesf ~ Elections_2016$nonwhite_pcts)

#Question 3: Is there a correlation between the proportion of the total number of votes per county and the percentage of females per county?
#plot of quantitative response variable and one of quantitative explanatory variables
plot(Elections_2016$female_pctn, Elections_2016$total_votesf, main = "Total Number of Votes vs Percentage of Females (Per County)", xlab = "Percentage of Females per County", ylab = "Total Number of Votes Per County")

#CONDUCTING TESTS

#RELATIONSHIP 1
#H0: ruralurban_ccn and total_votesx are independent
#Ha: ruralurban_ccn and total_votesx are dependent
#checking assumptions: data is randomly collected, observations are independent, cell counts greater than 5
Elections_2016$ruralurban_ccn
summary(Elections_2016$ruralurban_ccn)
summary(Elections_2016$total_votesx)
Elections_2016$ruralurban_ccn
table1 <- table(Elections_2016$total_votesx, Elections_2016$ruralurban_ccn)
table1
table2 <- prop.table(table1, margin = 2)
table2
#PLOT
barplot(table2, beside = T, legend.text = T, xlab = "Rural Urban Continuum Codes", ylab = "Proportion of votes in county")

```
test <- chisq.test(table1, correct = F)
test$expected
chisq.test(table1, correct = F)



#RELATIONSHIP 2
#H0: The difference in proportion of votes for the counties that are majority nonwhite and majority white is 0
#Ha: the difference in proportions of votes for the counties that are majority nonqhite and majority white is not 0
#total_votesf and nonwhite_pcts
#PLOT
boxplot(Elections_2016$total_votesf ~ Elections_2016$nonwhite_pcts, xlab = "County Racial Makeup", ylab = "Proportion of Votes in County")
#appropriate test is 2 sample means test
t.test(Elections_2016$total_votesf ~ Elections_2016$nonwhite_pcts, var.equal = FALSE)



#RELATIONSHIP 3
#quantitative
#linear regression
#H0: there is no correlation between total_votesf and female_pctn
#Ha: there is a correlation between total_votesf and female_pctn
hist(Elections_2016$female_pctn)
hist(Elections_2016$total_votesf)
#PLOT
plot(Elections_2016$female_pctn, Elections_2016$total_votesf, xlab = "% of Females per County", ylab =  "Proportion of Votes in County")
cor.test(Elections_2016$female_pctn, Elections_2016$total_votesf)
#correlation: 0.1297839, WEAK CORRELATION, VERY CLOSE TO 0


#illustrating linear regression
female_totalvotes_lm <- lm(Elections_2016$total_votesf ~ Elections_2016$female_pctn)
#for every 1% increase in females per county, the proportion of the total number of votes increases by 323.30254
summary(female_totalvotes_lm)
plot(Elections_2016$female_pctn, Elections_2016$total_votesf, xlab = "% of Females per County", ylab =  "Proportion of Total # of Votes Per County")
#adds regression line estimated from m1 to scatterplot
abline(female_totalvotes_lm)
```

```
#get confidence intervals for B0 and B1
confint(female_totalvotes_lm)


#STATS
install.packages("Rmisc")
library(Rmisc)

#TABLE 1

dim(Elections_2016)
summary(Elections_2016)

summary(table(Elections_2016$female_pctn, Elections_2016$nonwhite_pcts,
Elections_2016$ruralurban_ccn))
summary(Elections_2016$total_votesx)
#female_pctn
summarySE(data = Elections_2016, measurevar = "female_pctn", na.rm = T)
summarySE(data = Elections_2016, measurevar = "female_pctn", groupvars = "total_votesx", na.rm = T)

#nonwhite_pcts
summary(Elections_2016$nonwhite_pcts)

#ruralurban_ccn
summary(Elections_2016$ruralurban_ccn)

#total_votesf
summary(Elections_2016$total_votesf)
summarySE(data = Elections_2016, measurevar = "total_votesf", na.rm = T)
summarySE(data = Elections_2016, measurevar = "total_votesf", groupvars = "total_votesx", na.rm = T)


Elections_2016$total_votesx[Elections_2016$total_votesf <= 0.000317] <- "Less votes in county than
national county average"
summary(Elections_2016$total_votesx)

#RESEARCH QUESTION

#Question 1: Is the population of a county associated with whether a county has more or less votes than
the national county average number of votes?
#plot of dichotomous response variable and one of categorical explanatory variables
```

```
boxplot(Elections_2016$ruralurban_ccn ~ Elections_2016$total_votesx, main = "County Population vs Total Number of Votes per County  ", xlab = "Total Number of Votes per County", ylab = "Rural-Urban Continuum Codes")

#Question 2: Does the average proportion of the total number of votes for counties that are majority white differ from counties that are majority nonwhite?
#plot of quantitative response variable and a categorical explanatory variable
boxplot(Elections_2016$total_votesf ~ Elections_2016$nonwhite_pcts)


#Question 3: Is there a correlation between the  proportion of the total number of votes per county and the percentage of females per county?
#plot of quantitative response variable and one of quantitative explanatory variables
plot(Elections_2016$female_pctn, Elections_2016$total_votesf, main = "Total Number of Votes vs Percentage of Females (Per County)", xlab = "Percentage of Females per County", ylab =  "Total Number of Votes Per County")


#CONDUCTING TESTS


#RELATIONSHIP 1
#H0: ruralurban_ccn and total_votesx are independent
#Ha: ruralurban_ccn and total_votesx are dependent
#checking assumptions: data is randomly collected, observations are independent, cell counts greater than 5
Elections_2016$ruralurban_ccn
summary(Elections_2016$ruralurban_ccn)
summary(Elections_2016$total_votesx)
Elections_2016$ruralurban_ccn
table1 <- table(Elections_2016$total_votesx, Elections_2016$ruralurban_ccn)
table1
table2 <- prop.table(table1, margin = 2)
table2
#PLOT
barplot(table2, beside = T, legend.text = T, xlab = "rural urban continuum codes", ylab = "Proportion of votes in county to total # votes")
test <- chisq.test(table1, correct = F)
test$expected
chisq.test(table1, correct = F)
```

```
#RELATIONSHIP 2
#H0: The difference in proportion of votes for the counties that are majority nonwhite and majority white is 0
#Ha: the difference in proportions of votes for the counties that are majority nonqhite and majority white is not 0
#total_votesf and nonwhite_pcts
#PLOT
boxplot(Elections_2016$total_votesf ~ Elections_2016$nonwhite_pcts, xlab = "County Racial Makeup", ylab = "Proportion of votes in county to total # votes")
#appropriate test is 2 sample means test
t.test(Elections_2016$total_votesf ~ Elections_2016$nonwhite_pcts, var.equal = FALSE)



#RELATIONSHIP 3
#quantitative
#linear regression
#H0: there is no correlation between total_votesf and female_pctn
#Ha: there is a correlation between total_votesf and female_pctn
hist(Elections_2016$female_pctn)
hist(Elections_2016$total_votesf)
#PLOT
plot(Elections_2016$female_pctn, Elections_2016$total_votesf, xlab = "% of Females per County", ylab =  "Proportion of Total # of Votes Per County")
cor.test(Elections_2016$female_pctn, Elections_2016$total_votesf)
#correlation: 0.1297839, WEAK CORRELATION, VERY CLOSE TO 0


#illustrating linear regression
female_totalvotes_lm <- lm(Elections_2016$total_votesf ~ Elections_2016$female_pctn)
#for every 1% increase in females per county, the proportion of the total number of votes increases by 323.30254
summary(female_totalvotes_lm)
plot(Elections_2016$female_pctn, Elections_2016$total_votesf, xlab = "% of Females per County", ylab =  "Proportion of Total # of Votes Per County")
#adds regression line estimated from m1 to scatterplot
abline(female_totalvotes_lm)
#get confidence intervals for B0 and B1
confint(female_totalvotes_lm)


#STATS
install.packages("Rmisc")
```

```
library(Rmisc)

#TABLE 1

dim(Elections_2016)
summary(Elections_2016)

summary(table(Elections_2016$female_pctn, Elections_2016$nonwhite_pcts,
Elections_2016$ruralurban_ccn))

#female_pctn
summarySE(data = Elections_2016, measurevar = "female_pctn", na.rm = T)
summarySE(data = Elections_2016, measurevar = "female_pctn", groupvars = "total_votesx", na.rm = T)

#nonwhite_pcts
summary(Elections_2016$nonwhite_pcts)

#ruralurban_ccn
summary(Elections_2016$ruralurban_ccn)

#total_votesf
summary(Elections_2016$total_votesf)
summarySE(data = Elections_2016, measurevar = "total_votesf", na.rm = T)
summarySE(data = Elections_2016, measurevar = "total_votesf", groupvars = "total_votesx", na.rm = T)
```

Works Cited

Gould, Skye. “Americans Beat One Voter Turnout Record - Here's How 2016 Compares with Past Elections.” *Business Insider*, Business Insider, 21 Dec. 2016, www.businessinsider.com/trump-voter-turnout-records-history-obama-clinton-2016-11