Final Project: COVID19 and the 2020 Election

Author: Lukas Corey

Discussants:

538 prediction data: https://github.com/fivethirtyeight/data/blob/master/polls/README.md

County Covid Data: https://raw.githubusercontent.com/nytimes/covid-19-data/master/us-counties.csv

2016 election data https://public.opendatasoft.com/explore/dataset/usa-2016-presidential-election-by-county/information/?disjunctive.state

Census population data https://www.census.gov/data/tables/time-series/demo/popest/2010s-counties-total.html#par_textimage

2020 election data https://www.kaggle.com/unanimad/us-election-2020 https://www.kaggle.com/unanimad/us-election-2020?select=president_county_candidate.csv

Nate Silver on the Election and COVID: https://www.youtube.com/watch?v=O7bWIvQSvgk&list=UUXKjhxsfFQUqlNVQzLVnpEA

Plot Grid Info: https://www.rdocumentation.org/packages/cowplot/versions/1.1.0/topics/plot_grid

2d Density Plot w ggplot2: https://www.r-graph-gallery.com/2d-density-plot-with-ggplot2.html

```
library(dplyr, warn.conflicts = FALSE)
library(latex2exp)
library(ggplot2)
# for choropleth maps
library('maps')
library(viridis)
```

```
## Loading required package: viridisLite
```

```
# for plot grid
library(cowplot)
```

Introduction

COVID was central to the 2020 presidential election, from shaping how it was conducted to when votes from different population segments were counted to being a central issue for many voters. Especially with many still clinging onto claims of voter fraud and with the President saying that inaccurate predictions are a form of voter suppression, it is worth asking a couple questions about the polling data and how COVID affected the election. The answers to these questions are important for thinking about the conduct of coming elections, presidential and otherwise, and the status of the two major political parties in America.

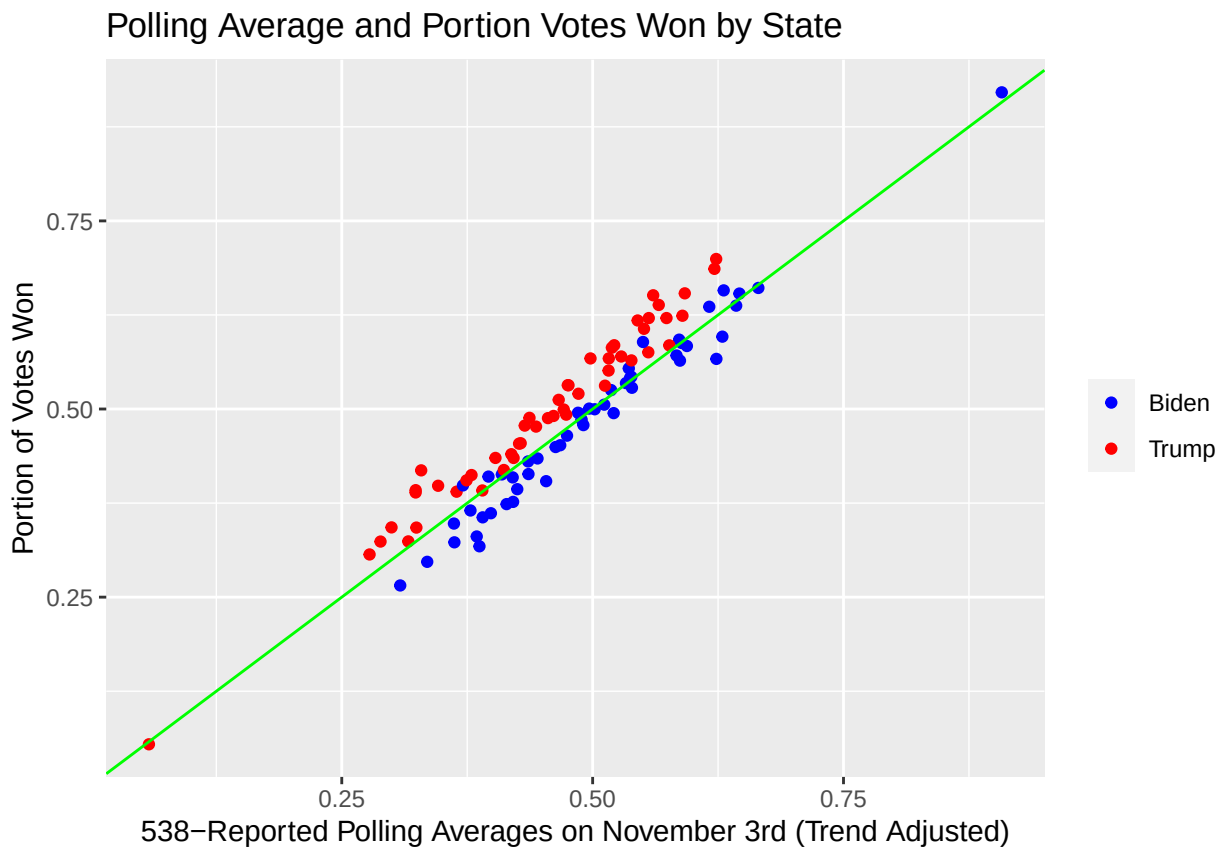
1. Why were polls significantly off this year? Although polling averages from 538 guessed the winner of every state but two correctly, there was still significant deviation from recent polls in the actual results. Siena College reported an 11 point lead for Biden in Wisconsin just days before the election (https://scri.siena.edu/2020/11/01/the-new-york-times-siena-college-battleground-polls-arizona-florida-pennsylvania-wisconsin/), but he won the state by only around .7 percent. Were the polls systematically underestimating Trump, and as Nate Silver suggested might be a reason, could this be at least partially due to COVID? ("If people are changing their living patterns around the pandemic, that might change how they are responding to polls too," said Nate Silver in the video linked above).

2. Third party votes (which I'll lump with votes for anyone other than the two major party candidates) are sometimes called throwaway votes by those concerned about the outcome of the election. Given how important people think the current challenges facing the country are (COVID, the economy, racial justice) and how divisive politics lead to polarized opinions on candidates, were third party votes much different this year than last election? Were they affected by COVID in some way? As some pollsters have suggested, is the Biden win largely attributable to an unequal allocation of third-party votes to Biden over Trump relative to 2016?

Data links are above in the discussants section, but voting data is from opendatasoft (2016) and unanimad on Kaggle (2020). Election predictions are from 538, COVID data is from the NYT, and population data is from the Census Bureau. The IZA Institute of Labor Economics has nice analysis of "The COVID-19 Pandemic and the 2020 U.S. Presidential Election" here: http://ftp.iza.org/dp13862.pdf. They look specifically at the effect of COVID on Trump's support, while I focus on other features of election data, but our conclusions are consistent with each other.

DATA VISUALIZATION

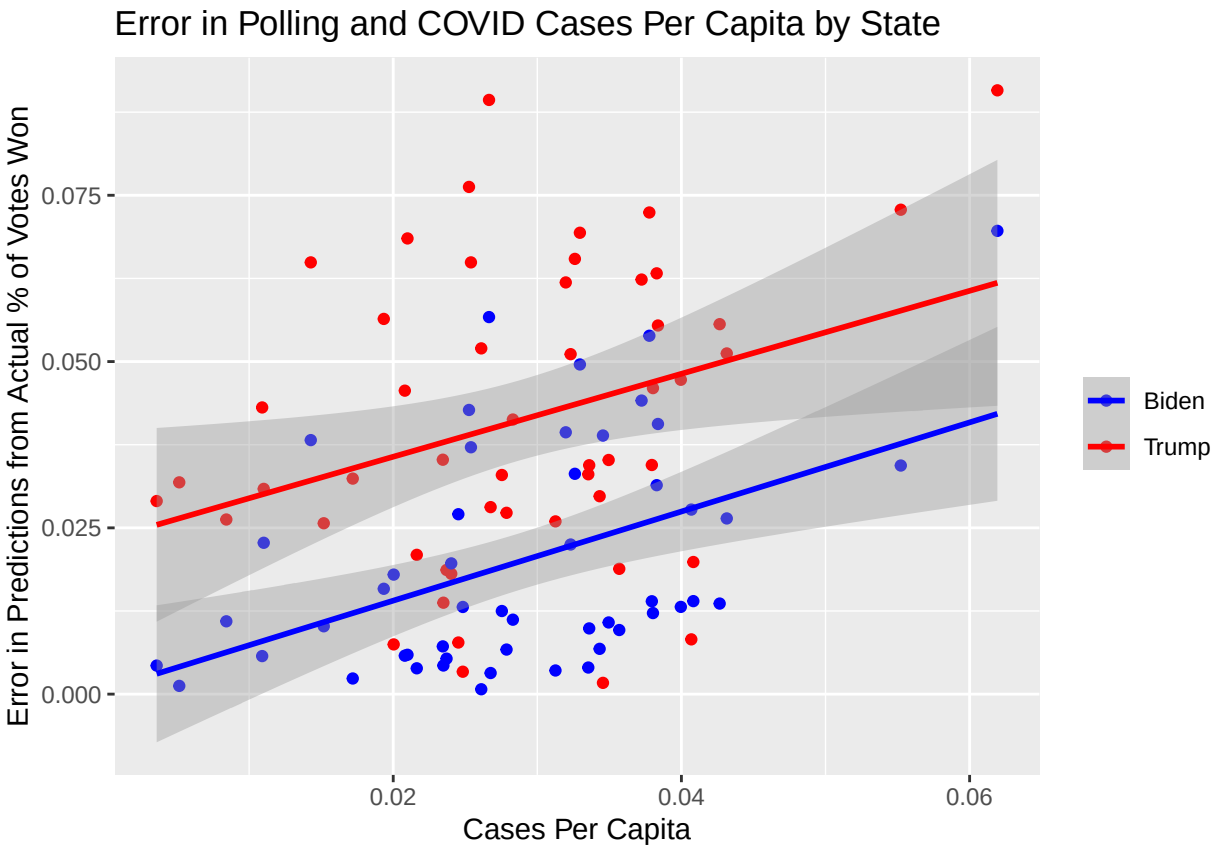
```
# See appendix A (Getting COVID Data By State) and B (538
# Predictions and Actual Election Results by State) for data
# wrangling here

ggplot(data = predictions_and_actuals, aes(pct_trend_adjusted,
    percent_won, color = cand_marker)) + geom_point() + geom_abline(slope = 1,
    intercept = 0, color = "green") + scale_color_manual(values = c("blue",
    "red")) + ggtitle("Polling Average and Portion Votes Won by State") +
    xlab("538-Reported Polling Averages on November 3rd (Trend Adjusted)") +
    ylab("Portion of Votes Won") + theme(legend.title = element_blank())
```



The graph above shows the systematic underestimation of Trump support by polls across states. The line in the middle is x=y, or where the polling averages exactly equal the outcome in a state. Trump fairly consistently won a higher portion of votes by state than he was predicted to, as is clear from the red points being above the line. Joe Biden generally underperformed, especially consistently in states where he wasn't projected to win. Many pollsters argued going into this election that Biden's lead was fundamentally different from Hillary Clinton's and that the polling errors from 2016 wouldn't return. Others were more realistic, but here we see a fairly similar underprediction of Trump's support. This is sometimes attributed to the "shy trump voter" hypothesis, but most who look carefully find no evidence for this explanation, instead arguing that weighing polls by demographics and regions is the issue.

```
# Data wrangling for this and all of the choropleth maps in
# Appendix C (Choropleth Maps for Covid Cases per Capita and
# Election Polling Error by State)

ggplot(data = error_and_covid, aes(cases_per_capita, error, color = cand_marker)) +
    geom_point() + geom_smooth(method = lm) + scale_color_manual(values = c("blue",
    "red")) + ggtitle("Error in Polling and COVID Cases Per Capita by State") +
    xlab("Cases Per Capita") + ylab("Error in Predictions from Actual % of Votes Won") +
    theme(legend.title = element_blank())
```

```
## `geom_smooth()` using formula 'y ~ x'
```



The above plot shows a statistically significant (see the modeling and analysis below) positive correlation between the number of Covid Cases per capita in a state and the error of the polling data (taken from 538's trend adjusted polling averages) when compared to actual election data for both Trump and Biden. The most dramatic examples of this (which are easier to visualize on the choropleth maps in appendix C) are North Dakota, Idaho, and New York, as well as some of the midwestern states. This suggests that there might be something to Nate Silver's claim that COVID caused polling error in some way. Further analysis of Silver's claim that this is due to "lifestyle changes with the pandemic" would require looking more specifically at lockdown measures by state, but this initial correlation between error and covid prevalence is in agreement. There are obviously many other potential explanations related to the type of locations with high covid prevalence (especially because I selected a single timepoint of Nov 3 to take covid data from).

DATA MODELING AND ANALYSIS

```
# Test for Correlation

# Get observed statistics
observed_correlation_trump <- cor(trump_error$cases_per_capita, trump_error$error)
observed_correlation_biden <- cor(biden_error$cases_per_capita, biden_error$error)

# Create null distributions
null_dist_trump <- NULL
null_dist_biden <- NULL
for (i in 1:10000){
  null_dist_trump[i] <- cor(trump_error$cases_per_capita, sample(trump_error$error))
```

```
  null_dist_biden[i] <- cor(biden_error$cases_per_capita, sample(biden_error$error))
}

# Plot distributions -- MOVED TO APPENDIX D
# plot_grid(biden_plot, trump_plot, labels = "AUTO")

# Manually calculate the p values
pval_trump <- (sum(null_dist_trump >= observed_correlation_trump) +
                 sum(null_dist_trump <= -observed_correlation_trump))/length(null_dist_trump)
pval_biden <- (sum(null_dist_biden >= observed_correlation_biden) +
                 sum(null_dist_biden <= -observed_correlation_biden))/length(null_dist_biden)

# Confidence intervals for correlations
interval_trump <- cor.test(trump_error$cases_per_capita, trump_error$error)$conf.int[1:2]
interval_biden <- cor.test(biden_error$cases_per_capita, biden_error$error)$conf.int[1:2]

print(paste("The correlation between Biden vote percent error and covid cases per capita has a confiden
```

```
## [1] "The correlation between Biden vote percent error and covid cases per capita has a confidence in
```

"The correlation between Biden vote percent error and covid cases per capita has a confidence interval of 0.209615674793535 - 0.651880666776459 and for Trump vote percent error the 95 % CI is 0.0506577338993007 - 0.548682232828706 with p values of 9e-04 and 0.0203 respectively." Our Null hypothesis is that there is no correlation between the number of COVID cases per capita and the error in polling for Biden/Trump. $H_0 : \rho = 0$ Our Alternative hypothesis is that there is a positive correlation between the number of COVID cases per capita and the error in polling for Biden/Trump. $H_a : \rho > 0$ With a significance level of $\alpha = .05$, we can reject the null hypothesis and are forced to accept that there is a positive correlation for both, with p values of 9e-04 for Biden's error and .0203 for Trumps (see above). The confidence 95 percent confidence intervals for the correlations are above, and both exclude 0 as we would expect.

DATA WRANGLING

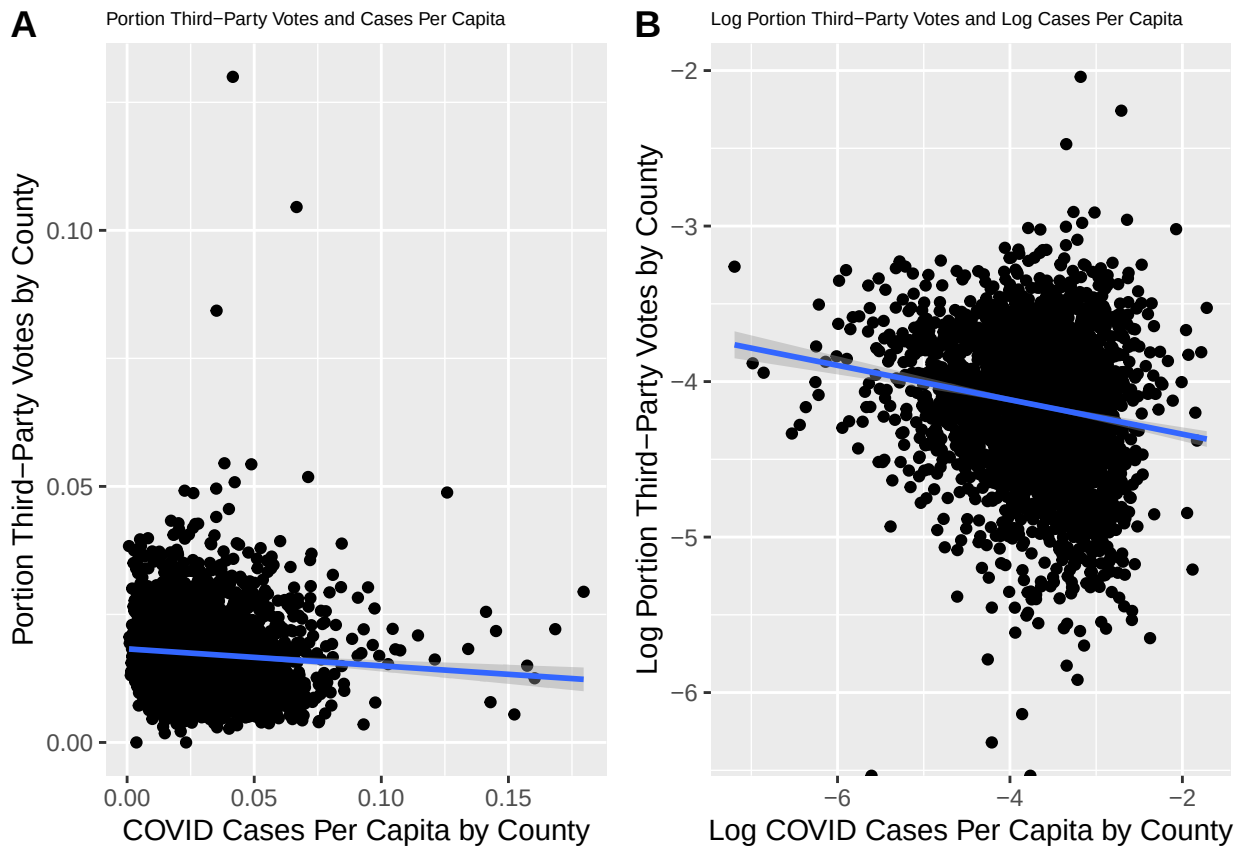
```
# Much more in Appendix Goal here is to get cases per capita
# and votes that went to a non-major-party candidate to graph
# get covid cases by county from the day before the election
by_county <- county_covid_cases %>% dplyr::filter(date == "2020-11-03") %>%
    # format state to match other data frame
dplyr::mutate(state = tolower(state))

# get population data from census dataset and select relevant
# columns
county_pop <- read.csv("co-est2019-alldata.csv") %>% dplyr::select(POPESTIMATE2019,
    STNAME, CTYNAME) %>% dplyr::mutate(state = STNAME) %>% dplyr::mutate(county = CTYNAME)

# get votes by county
county_votes <- read.csv("president_county_candidate.csv")

# fix incorrect letter representation in census data that
# causes errors
county_pop[which(county_pop$CTYNAME == "Do\xf1a Ana County"),
```

```
    ]$CTYNAME <- "Doña Ana County"

# join population data and voting data
county_pop_plus_votes <- left_join(county_pop, county_votes,
    by = c("state", "county"))

# modify county names to match in all data frames by removing
# the word 'County' if it's at the end for example, King
# County -> King cannot just add county to all of those
# without it, because some like 'Disitrict of Columbia' don't
# have county at the end
county_name_minus_county <- NULL
current_names <- county_pop_plus_votes$CTYNAME
for (i in 1:length(current_names)) {
    if (substring(c(current_names[i]), nchar(current_names[i]) -
        5, nchar(current_names[i])) == "County") {
        county_name_minus_county[i] <- substring(c(current_names[i]),
            1, nchar(current_names[i]) - 7)
    } else {
        county_name_minus_county[i] <- current_names[i]
    }
}

# replace column with fixed names
county_pop_plus_votes$county <- county_name_minus_county

# format state as lowercase for joining in population +
# voting data data frame.
county_pop_plus_votes <- county_pop_plus_votes %>% dplyr::mutate(state = tolower(state))

# join pop data, covid data, and voting data and calculate
# covid cases per capita as well as third party votes
county_cases_per_capita_plus_votes <- na.omit(left_join(by_county,
    county_pop_plus_votes, by = c("state", "county")) %>% dplyr::mutate(cases_per_capita = cases/POPESTI
    # label all non-major party candidates as other in a new
# column
dplyr::mutate(cand_marker = ifelse(candidate == "Joe Biden",
    "Biden", ifelse(candidate == "Donald Trump", "Trump", "other"))) %>%
    # get total votes for each county
group_by(state, county) %>% dplyr::mutate(county_total_votes = sum(total_votes)) %>%
    # get total votes for each candidate and for all non-major
# party candidates
group_by(state, county, cand_marker) %>% dplyr::mutate(cand_marker_total_votes = sum(total_votes)) %>%
    # remove third party candidate names and repeated rows
distinct(state, county, cand_marker, .keep_all = TRUE) %>% dplyr::mutate(percent_votes_won = cand_marke
    dplyr::select(state, county, total_votes, cases_per_capita,
        cand_marker, percent_votes_won)

# get data for third party candidates and log transform for
# better visualization
others <- na.omit(county_cases_per_capita_plus_votes %>% dplyr::filter(cand_marker ==
    "other")) %>% dplyr::mutate(log_percent_votes_won = log(percent_votes_won)) %>%
    dplyr::mutate(log_cases_per_capita = log(cases_per_capita))
```

DATA VISUALIZATION and MODELING

```
# Data Wrangling in Appendix E

# Plot effect of covid cases on portion of third party votes
nonlogplot <- ggplot(others, aes(cases_per_capita, percent_votes_won)) +
    geom_point() + geom_smooth(method = lm) + ggtitle("Portion Third-Party Votes and Cases Per Capita") +
    xlab("COVID Cases Per Capita by County") + ylab("Portion Third-Party Votes by County") +
    theme(plot.title = element_text(size = 7))
logplot <- ggplot(others, aes(log_cases_per_capita, log_percent_votes_won)) +
    geom_point() + geom_smooth(method = lm) + ggtitle("Log Portion Third-Party Votes and Log Cases Per C
    xlab("Log COVID Cases Per Capita by County") + ylab("Log Portion Third-Party Votes by County") +
    theme(plot.title = element_text(size = 7))

# Plot side by side
plot_grid(nonlogplot, logplot, labels = "AUTO")
```

```
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
```




```
# correlation test for percent of third party votes and cases
# per capita by county
(thirdpartyvotes_and_casespercapita <- cor.test(others$cases_per_capita,
    others$percent_votes_won))
```

```
##
##  Pearson's product-moment correlation
##
## data:  others$cases_per_capita and others$percent_votes_won
## t = -4.2211, df = 2970, p-value = 2.504e-05
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.11286486 -0.04138372
## sample estimates:
##         cor
## -0.07722353
```

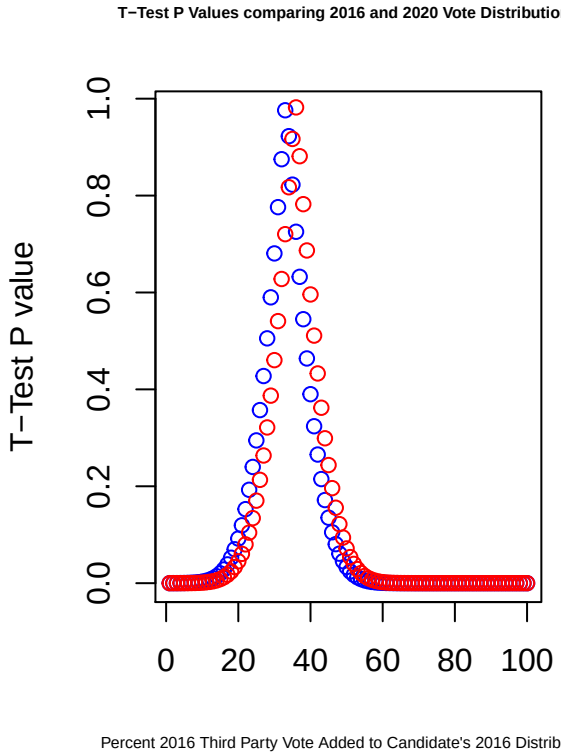
```
# Model Analyzed in Appendix F (including residuals and cooks
# distances) thirdparty_covid_model <-
# lm(cases_per_capita~percent_votes_won, data = others)
```

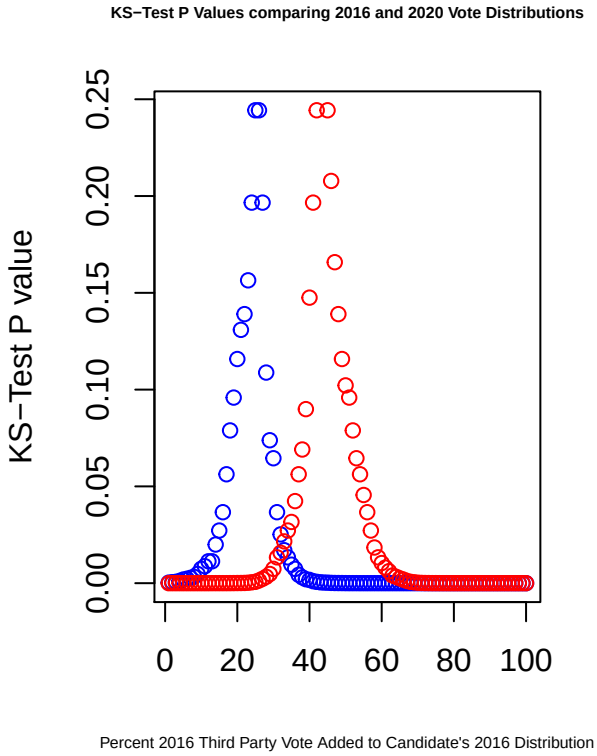
Here we can see a statistically significant (see analysis in Appendix F) decrease in the portion of votes going to a third party candidate with an increase in number of covid cases per county. There are two obvious potential explanations for this: that counties with higher number of covid cases and more likely to be republican (or democratic) and are more attached to their candidate or (my preferred explanation) that covid is a very polarizing issue and the more it affects someone, the more likely they are to have a strong opinion on which major party candidate would address it better and avoid voting third party. In other words, if COVID is prevalent in your county, you become a single issue voter and will avoid "wasting" your vote on a non-major-party candidate.

DATA MODELING

```
# Analysis of the Basic change (decrease) in
# non-major-candidate vote in Appendix G

# Data Wrangling for this in Appendix H

# The idea here is to see whether Biden's vote percentages
# can be modeled accurately by taking Hillary's in 2016 and
# adding some portion of the non-major-candidate vote I add
# 1->100 percent of the third party vote to Hillary's 2016
# county percents and then t test or ks test with Bidens 2020
# percentages by county.  I did the same with Trumps numbers,
# and plotted these together. See analysis below.

par(mfrow = c(1, 2))

# T test comparing biden 2020 percent won with (hillary 2016
# percent won by county plus __% of third party vote)
ttestvals <- plot(seq(1, 100, 1), ttest_p_values_biden, col = "blue",
    main = list("T-Test P Values comparing 2016 and 2020 Vote Distributions",
        cex = 0.5), xlab = list("Percent 2016 Third Party Vote Added to Candidate's 2016 Distribution",
        cex = 0.5), ylab = "T-Test P value")
points(ttest_p_values_trump, col = "red")

# Kolmogorov-Smirnov test comparing biden 2020 percent won
```

```r
# with (hillary 2016 percent won by county plus __% of third
# party vote)
kstestvals <- plot(seq(1, 100, 1), kstest_p_values_biden, col = "blue",
    main = list("KS-Test P Values comparing 2016 and 2020 Vote Distributions",
        cex = 0.5), xlab = list("Percent 2016 Third Party Vote Added to Candidate's 2016 Distribution",
        cex = 0.5), ylab = "KS-Test P value")
# Add in Trump values
points(kstest_p_values_trump, col = "red")
```





Political pundits and news media said often during the election and after that it seemed as if Joe Biden was repeating Hillary Clinton numbers with the addition of the third party/independent/non-major-party candidate vote (which I use interchandeably here). This is disputed by the above graphs.

The graph on the left suggests that Biden's 2020 vote percentages by county as best modeled by adding about 33 percent of the independent vote from 2016 to Clinton's 2016 numbers (because this is where the p value is the highest and there's the least evidence to reject that these distributions came from different data sets), but that the same is true for Trump's 2020 and 2016 numbers. This suggests that a better characterization of the change in third party vote is a split (see the decrease in people voting third party in appendix G) between Trump and Biden. However, these additional third party votes to Biden came in areas that helped him win. I would guess that additional third party votes for Trump came from areas where his base was loud and prominent for four years (and where it didn't help him to get additional support). Although the election was eventually a win for Biden, it's important to remember that Trump actually increased his percentage of the popular vote won nationally and in many states.

The graph on the right uses the Kolmogorov Smirnov test, which looks for maximum difference between two distributions rather than primarily being based on the median value. With it's p value maxing out around .25, it suggests there is no place where the distributions are not noticeably different for both Biden and Trump.

It's inaccurate to characterize Biden's win as a simple across-the-board shifting of the independent vote to him both because this didn't really happen and because the by-county percent maps were fundamentally different in this election compared to the last, whether because of shifting demographics or differential appeal of Biden and Clinton or something else entirely.

Conclusion

Because this election was a victory for Biden, and ultimately a fairly convincing one both in the electoral college and the popular vote, I think it's easy to ignore important lessons from the election and analyze it incorrectly. 1. There was significant error in the polling, and it underestimated republican/Trump support once again ("Polling averages and Portion Votes won by state" graph on page 3). 2. This error in polling was compounded by, and correlated with, COVID ("Error in polling and COVID cases per capita by state" graph page 4, see also choropleth maps in Appendix C). Having accurate polling data is always difficult, but it will continue to increase difficulty in the future if we see sustained lifestyle changes regarding working from home or moving to new locations post-COVID (or if whatever other explanation accounts for the error correlated with COVID persists, for example mail in voting remaining prevalent may shift populations and force models to adjust). 3. While there was a decrease in non-major-party candidate voting (2d graphs of non-major-party-candidate votes in 2016 and 2020 in appendix G), it was not the shift to Biden we might assume. It was correlated with covid incidence ("Portion Third-Party Votes and Cases Per Capita by County" graph page 7) which suggests it may be related to peoples' opinions on COVID, but likely relatively evenly split on a national level between candidates (comparison of 2020 voting distributions and 2016 distributions plus third party votes on page 9). This suggests Biden wasn't as much of a favorite on COVID as we might assume. This is further suggested by the data in appendix I showing a negative correlation between Biden improvement on Clinton's performance and covid incidence. These findings provide interesting insight into the status of the democratic and republican parties in the national eye, but also insight into the polarized state of every issue. COVID is an issue on which Trump should clearly underperform, but the split of third party voters and the loyalty of Trump's base suggest this wasn't a dominant effect. Democrats were hoping for a repudiation of Trump that perhaps wasn't as clear as it should've been, with tight margins in Wisconsin, Georgia, and Michigan handing Biden the victory. It's also worth mentioning that Trump actually increased his percent of votes nationally from 2016, gaining support, even though Biden won. Worthwhile future directions include investigating correlations between opinion polls and voting in areas affected by COVID to varying degrees. COVID is an issue many downplay and others focus on intensely–is this reflected in opinion polls and how people vote? How willing are people to believe something is an issue if it isn't in their neighborhood and affecting people they know? This comments on political strategy and how to sell new policy as well as the state of our democracy.

Reflection

One major challenge was formatting the data appropriately from different sources for joining data frames. County name feels like something that should be fairly standard, but when working to bring together census population data, election data from 2020, election data from 2016, and covid county data, I found that some included the word county while others didn't. Some included the state name after the county name separated by a comma. Sorting through all of this is a major reason I have a huge appendix with much dplyr data wrangling. With this, it's also tough to keep track of which data frames I'm working with and which have the data I want. I believe I did end up losing a significant number of counties in my joining, and it's tough to know whether and how much this affects the analysis. Another challenge was properly displaying all of the data in different ways and trying to incorporate good breadth of statistical methodology with a good breadth of data visualization methods and showing my data wrangling. Fitting this all into ten pages (a TA on piazza said we could have up to ten without losing points) is very tough. I looked a little bit at the Georgia data and how it shifted from 2016 to 2020, but did not end up including this here. I estimate that I spent ~15-20 hours working on this.

Appendix

Appendix A: Getting COVID Data By State

```
county_covid_cases <- read.csv('https://raw.githubusercontent.com/nytimes/covid-19-data/master/us-counti

state_pop <- read.csv("statepop.csv")

state_cases <- county_covid_cases %>%
  dplyr::filter(date == "2020-11-03") %>%
  dplyr::group_by(state) %>%
  dplyr::mutate(statesum = sum(cases), statedeaths = sum(deaths)) %>%
  dplyr::distinct(state, .keep_all = TRUE) %>%
  dplyr::mutate(state = tolower(state))

state_cases_per_capita <- na.omit(left_join(state_cases, state_pop, by = "state") %>%
  dplyr::mutate(cases_per_capita = statesum / pop_2019) %>%
  dplyr::select(state, statesum, statedeaths, pop_2019, cases_per_capita))
```

Appendix B: 538 Predictions and Actual Election Results by State

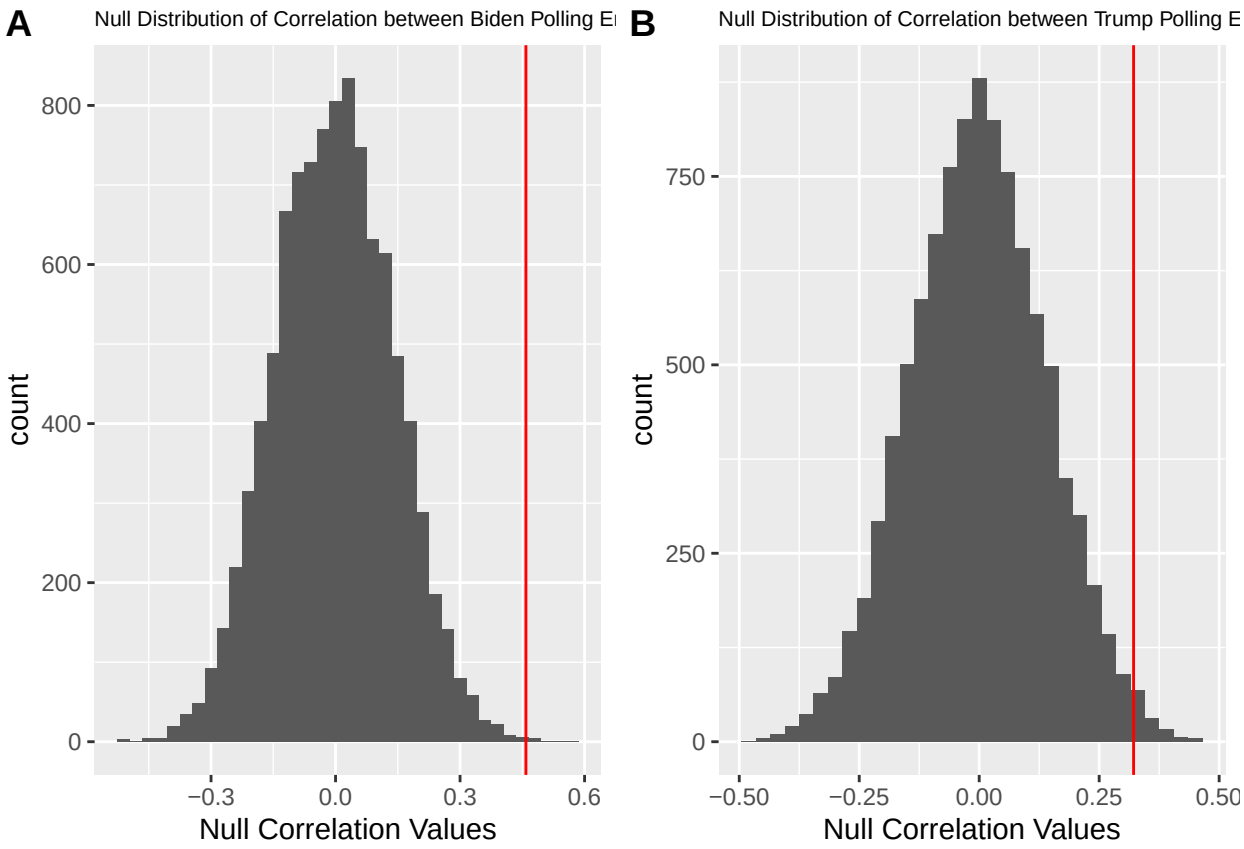
```
candidate_vote_percents_by_state <- read.csv("president_county_candidate.csv")  %>%
  dplyr::group_by(state) %>%
  dplyr::mutate(state_votes = sum(total_votes)) %>%
  dplyr::group_by(state, candidate) %>%
  dplyr::mutate(state_candidate_votes = sum(total_votes)) %>%
  dplyr::distinct(state, .keep_all = TRUE) %>%
  dplyr::ungroup() %>%
  dplyr::mutate(percent_won = state_candidate_votes/state_votes) %>%
  dplyr::mutate(cand_marker = ifelse(candidate == "Joe Biden", "Biden",
                                     ifelse(candidate == "Donald Trump", "Trump", "other"))) %>%
  dplyr::select(state, cand_marker, state_votes, state_candidate_votes, percent_won)


predictions <- read.csv("presidential_poll_averages_2020.csv") %>%
  dplyr::filter(candidate_name == "Joseph R. Biden Jr." | candidate_name == "Donald Trump") %>%
  dplyr::filter(modeldate == "11/3/2020") %>%
  dplyr::mutate(cand_marker = ifelse(candidate_name == "Joseph R. Biden Jr.", "Biden", ifelse(candidate_
  dplyr::select(state, cand_marker, pct_trend_adjusted) %>%
  dplyr::mutate(pct_trend_adjusted = as.numeric(pct_trend_adjusted) / 100)

predictions_and_actuals <- inner_join(predictions, candidate_vote_percents_by_state, by = c("cand_marke
```

Appendix C: Choropleth Maps for Covid Cases per Capita and Election Polling Error by State

```
predictions_and_actuals <- predictions_and_actuals %>%
  dplyr::mutate(error = abs(pct_trend_adjusted - (state_candidate_votes/state_votes))) %>%
  dplyr::arrange(error) %>%
  dplyr::mutate(state = tolower(state))

error_and_covid <- na.omit(inner_join(predictions_and_actuals, state_cases_per_capita, by = "state") %>%
  dplyr::select(error, state, cases_per_capita, cand_marker))

trump_error <- error_and_covid[which(error_and_covid$cand_marker == "Trump"),]

biden_error <-error_and_covid[which(error_and_covid$cand_marker == "Biden"),]

trump_error_map <- inner_join(x = (map_data("state")), y = trump_error, by = c("region" = "state")) %>%
biden_error_map <- inner_join(x = (map_data("state")), y = biden_error, by = c("region" = "state")) %>%


covid_map <- ggplot() + geom_polygon(data = trump_error_map, aes(x = long, y = lat, fill = cases_per_cap
  theme(axis.title.x=element_blank(),axis.text.x=element_blank(),axis.ticks.x=element_blank(),
        axis.title.y=element_blank(),axis.text.y=element_blank(),axis.ticks.y=element_blank(), legend.ti

trump_error_m <- ggplot() + geom_polygon(data = trump_error_map, aes(x = long, y = lat, fill = error, gr
  theme(axis.title.x=element_blank(),axis.text.x=element_blank(),axis.ticks.x=element_blank(),
        axis.title.y=element_blank(),axis.text.y=element_blank(),axis.ticks.y=element_blank())

biden_error_m <- ggplot() + geom_polygon(data = biden_error_map, aes(x = long, y = lat, fill = error, gr
  theme(axis.title.x=element_blank(),axis.text.x=element_blank(),axis.ticks.x=element_blank(),
        axis.title.y=element_blank(),axis.text.y=element_blank(),axis.ticks.y=element_blank())

# Plot all with cowplot plot_grid
plot_grid(covid_map, trump_error_m, biden_error_m, labels = "AUTO")
```

Appendix D: Histograms

```
biden_plot <- ggplot(as.data.frame(null_dist_biden), aes(null_dist_biden)) + geom_histogram(binwidth =
  ggtitle("Null Distribution of Correlation between Biden Polling Error and COVID Cases per Capita") + x
trump_plot <- ggplot(as.data.frame(null_dist_trump), aes(null_dist_trump)) + geom_histogram(binwidth =
  ggtitle("Null Distribution of Correlation between Trump Polling Error and COVID Cases per Capita") + x

plot_grid(biden_plot, trump_plot, labels = "AUTO")
```



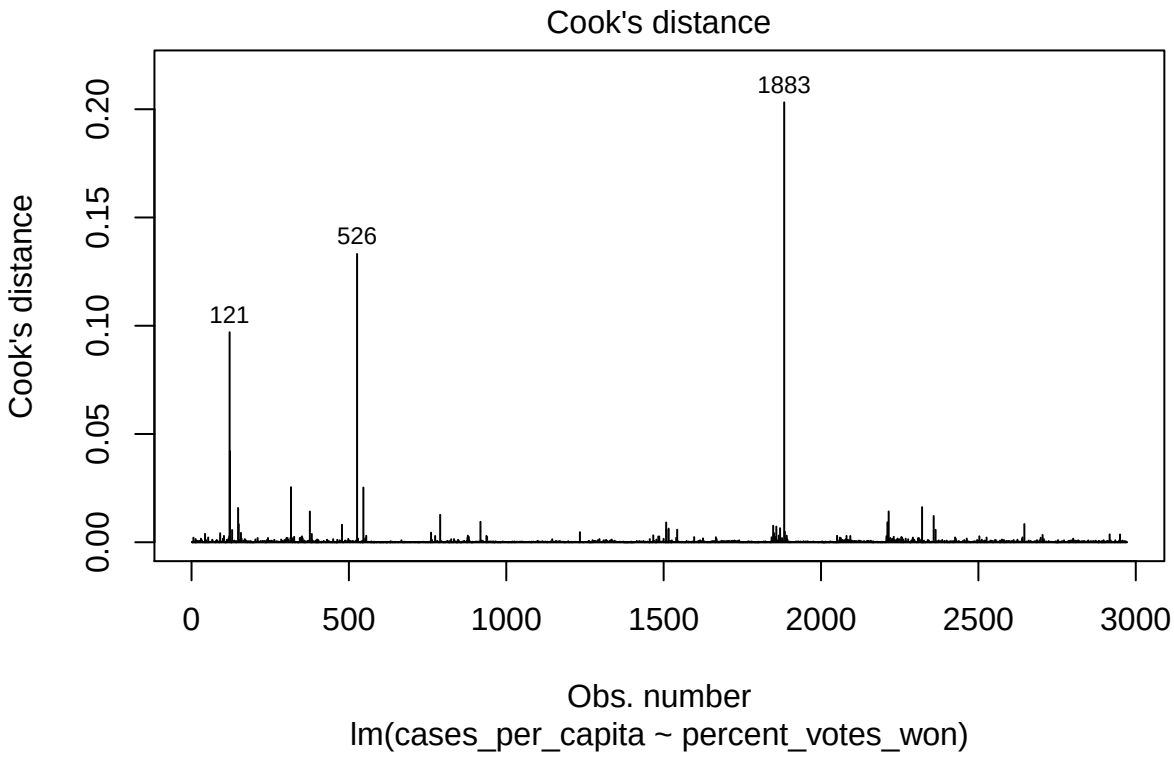
Appendix E: COVID Cases and Non-Major-Party Votes

```
by_county <- county_covid_cases %>%
  dplyr::filter(date == "2020-11-03") %>%
  dplyr::mutate(state = tolower(state))

county_pop <- read.csv("co-est2019-alldata.csv") %>%
  dplyr::select(POPESTIMATE2019, STNAME, CTYNAME) %>%
  dplyr::mutate(state = STNAME) %>%
  dplyr::mutate(county = CTYNAME)

county_votes <- read.csv("president_county_candidate.csv")

county_pop[which(county_pop$CTYNAME == "Do\xf1a Ana County"),]$CTYNAME <- "Doña Ana County"


county_pop_plus_votes <- left_join(county_pop, county_votes, by = c("state", "county"))

county_name_minus_county <- NULL
current_names <- county_pop_plus_votes$CTYNAME


for (i in 1:length(current_names)){
```

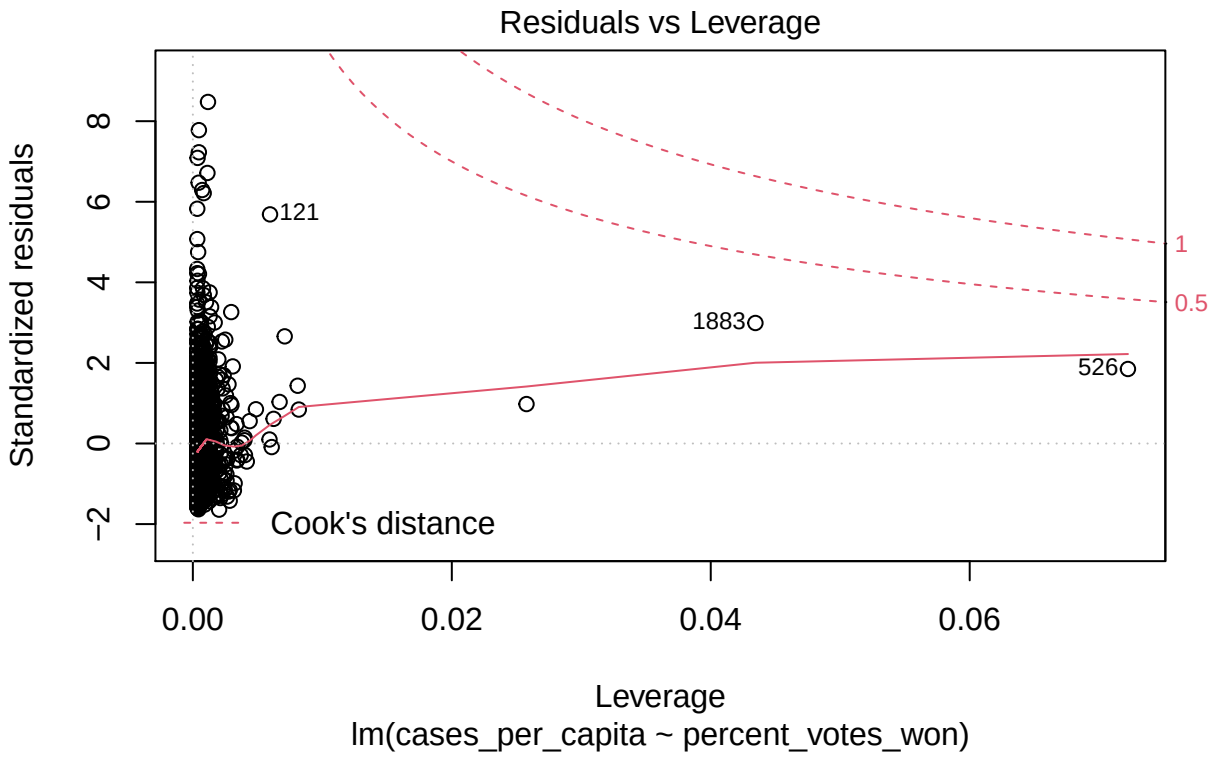
```
  if (substring(c(current_names[i]), nchar(current_names[i]) - 5,
                nchar(current_names[i])) == "County"){
    county_name_minus_county[i] <- substring(c(current_names[i]), 1,
                                             nchar(current_names[i]) - 7)}
  else{county_name_minus_county[i] <- current_names[i]}
}

county_pop_plus_votes$county <- county_name_minus_county

county_pop_plus_votes <- county_pop_plus_votes %>% dplyr::mutate(state = tolower(state))

county_cases_per_capita_plus_votes <- na.omit(left_join(by_county, county_pop_plus_votes,
                                          by = c("state", "county")) %>%
  dplyr::mutate(cases_per_capita = cases / POPESTIMATE2019)) %>%
  dplyr::mutate(cand_marker = ifelse(candidate == "Joe Biden", "Biden",
                                     ifelse(candidate == "Donald Trump", "Trump", "other"))) %>%
  group_by(state, county) %>%
  dplyr::mutate(county_total_votes = sum(total_votes)) %>%
  group_by(state, county, cand_marker) %>%
  dplyr::mutate(cand_marker_total_votes = sum(total_votes)) %>%
  distinct(state, county, cand_marker, .keep_all = TRUE) %>%
  dplyr::mutate(percent_votes_won = cand_marker_total_votes/county_total_votes) %>%
  dplyr::select(state, county, total_votes, cases_per_capita, cand_marker, percent_votes_won)

others <- na.omit(county_cases_per_capita_plus_votes %>% dplyr::filter(cand_marker == "other")) %>%
  dplyr::mutate(log_percent_votes_won = log(percent_votes_won)) %>%
  dplyr::mutate(log_cases_per_capita = log(cases_per_capita))
```

Appendix F: Third Party Vote Percent and COVID Incidence Model Analysis

```
thirdparty_covid_model <- lm(cases_per_capita~percent_votes_won, data = others)

plot(thirdparty_covid_model, 4)
```




```
plot(thirdparty_covid_model, 5)
```




```r
# North Dakota, Sioux many covid cases per pop
# others[1883,]
# Idaho, Camas population 1106
# others[526,]
# Arkansas, Lee
# Has ~1247 covid cases out of ~9000 people currently
# others[121,]

without_points <- others[-c(121, 526, 1883),]

summary(newmodel <- lm(cases_per_capita~percent_votes_won, data = without_points))
```

```
##
## Call:
## lm(formula = cases_per_capita ~ percent_votes_won, data = without_points)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.030580 -0.012397 -0.002883  0.008253  0.152787
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        0.0342568  0.0008424  40.667  < 2e-16 ***
## percent_votes_won -0.2533133  0.0451367  -5.612 2.18e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

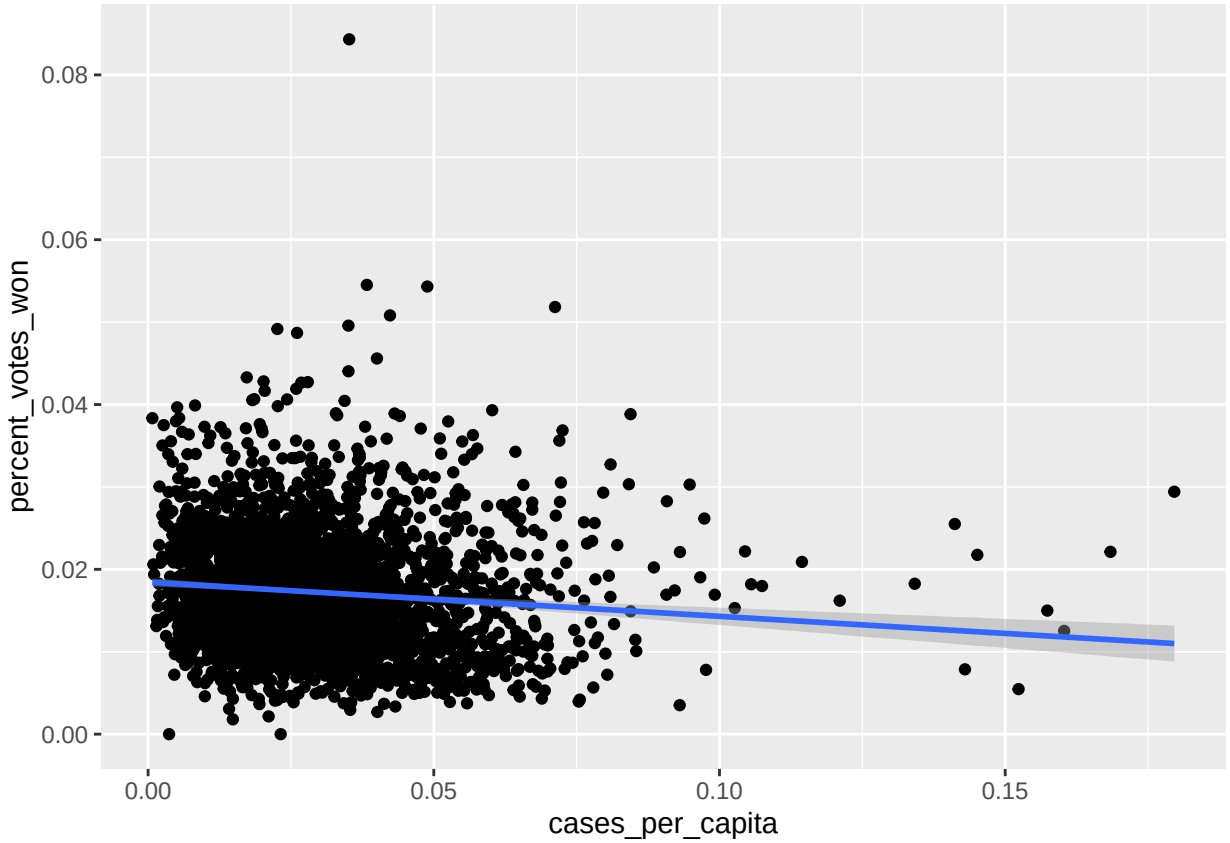
```
##
## Residual standard error: 0.01779 on 2967 degrees of freedom
## Multiple R-squared:  0.0105, Adjusted R-squared:  0.01017
## F-statistic:  31.5 on 1 and 2967 DF,  p-value: 2.183e-08
```

```
summary(thirdparty_covid_model)
```

```
##
## Call:
## lm(formula = cases_per_capita ~ percent_votes_won, data = others)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.029381 -0.012298 -0.002869  0.008376  0.151829
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        0.0330578  0.0008069  40.970  < 2e-16 ***
## percent_votes_won -0.1799814  0.0426384  -4.221  2.5e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.01792 on 2970 degrees of freedom
## Multiple R-squared:  0.005963,   Adjusted R-squared:  0.005629
## F-statistic: 17.82 on 1 and 2970 DF,  p-value: 2.504e-05
```

```
ggplot(without_points, aes(cases_per_capita, percent_votes_won)) + geom_point() + geom_smooth(method =
```

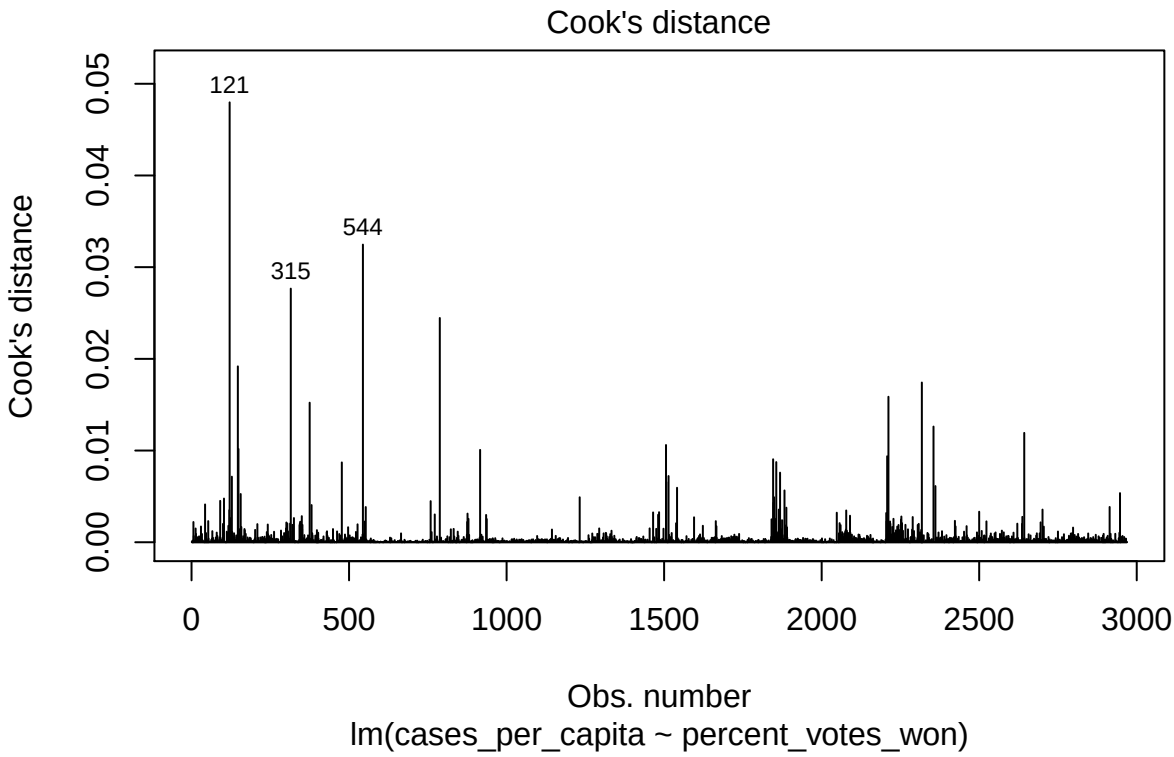
```
## `geom_smooth()` using formula 'y ~ x'
```



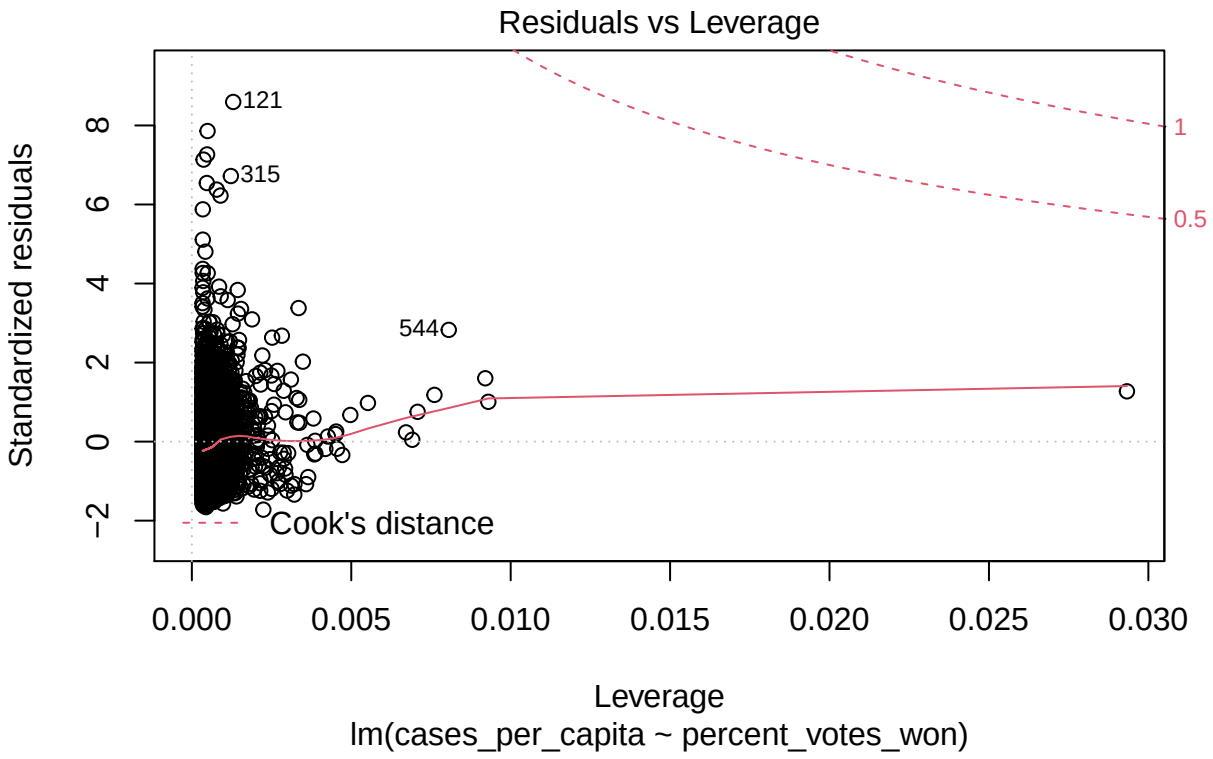

```
(cooks_limit <- mean(without_points$percent_votes_won) * 3)
```

```
## [1] 0.05161278
```

```
# no points with cooks distance above the new limit, and model is very similar (although cannot be dire
plot(newmodel, 4)
```



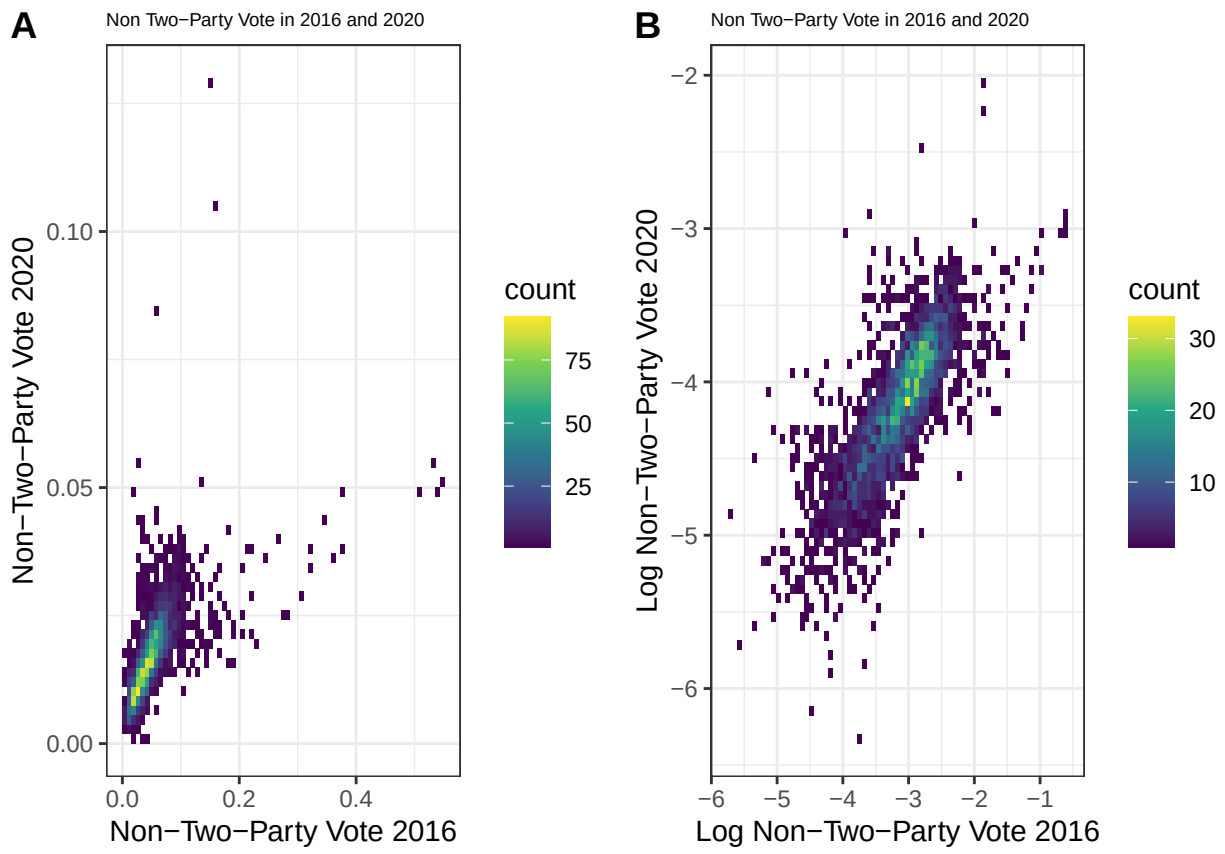

```
plot(newmodel, 5)
```



Appendix G: Change in Independent Vote Distribution from 2016 to 2020

```
# get 2016 data from election
prev_election_data <- read.csv("usa-2016-presidential-election-by-county.csv", sep=";") %>%
  dplyr::select(State, County, Votes, contains("Votes16."))

# non clinton or trump votes
other_votes <- prev_election_data %>%
  dplyr::select(County,
                (starts_with("Votes16") & (!contains("Clinton") & !contains("Trump")))) %>%
  replace(is.na(.), 0) %>%
  dplyr::mutate(other = rowSums(.[2:31])) %>%
  dplyr::select(County, other)

#join previous election data trump/clinton with summed third party data
prev_election_data <- na.omit(inner_join(prev_election_data %>%
  dplyr::select(State, County, Votes16.Clintonh, Votes16.Trumpd), other_votes, by = "County")) %>%
  dplyr::mutate(other_percent_won = other / (Votes16.Clintonh + Votes16.Trumpd))

#rename the counties for joining
county_names <- NULL
for (i in 1:length(prev_election_data$County)){
```

```
  minus_comma <- unlist(strsplit(prev_election_data$County[i], ","))[1]
  if (substring(c(minus_comma), nchar(minus_comma) - 5,
                nchar(minus_comma)) == "County"){
    county_names[i] <- substring(c(minus_comma), 1,
                                              nchar(minus_comma) - 7)
  }
  else{
    county_names[i] <- minus_comma
  }
}

# data frames, rename columns for clarity, and select appropriate columns
prev_election_data$county <- county_names
trump_twentytwenty_data <- county_cases_per_capita_plus_votes[county_cases_per_capita_plus_votes$cand_ma
biden_twentytwenty_data <- county_cases_per_capita_plus_votes[county_cases_per_capita_plus_votes$cand_ma
other_twentytwenty_data <- county_cases_per_capita_plus_votes[county_cases_per_capita_plus_votes$cand_ma
both_election_data <- na.omit(inner_join(inner_join(inner_join(prev_election_data %>%
                                          dplyr::mutate(state = tolower(State)),
                                        trump_twentytwenty_data,
                                        by = c("state", "county")),
                                        biden_twentytwenty_data,
                                        by = c("state", "county")),
                              other_twentytwenty_data, by = c("state", "county"))) %>%
  dplyr::mutate(other_2016 = other) %>%
  dplyr::mutate(other_percent_won_2016 = other_percent_won) %>%
  dplyr::mutate(trump_2020_votes = total_votes.x) %>%
  dplyr::mutate(biden_2020_votes = total_votes.y) %>%
  dplyr::mutate(other_2020_votes = total_votes) %>%
  dplyr::mutate(other_2020_percent = percent_votes_won) %>%
  dplyr::select(state, county, other_2016, other_percent_won_2016, trump_2020_votes, biden_2020_votes,
                Votes16.Clintonh, Votes16.Trumpd, other_2020_votes, cases_per_capita)

both_election_data$logother2016 <- log(both_election_data$other_percent_won_2016)
both_election_data$logother2020 <- log(both_election_data$other_2020_percent)

# plot relationship between other vote in 2016 and other votes in 2020
nonlogplot <- ggplot(both_election_data, aes(x=other_percent_won_2016, y=other_2020_percent)) +
  geom_bin2d(bins = 70) +
  scale_fill_continuous(type = "viridis") +
  theme_bw() + ggtitle("Non Two-Party Vote in 2016 and 2020")+ xlab("Non-Two-Party Vote 2016") + ylab("

logplot <- ggplot(both_election_data, aes(x=logother2016, y=logother2020)) +
  geom_bin2d(bins = 70) +
  scale_fill_continuous(type = "viridis") +
  theme_bw() + ggtitle("Non Two-Party Vote in 2016 and 2020") + xlab("Log Non-Two-Party Vote 2016") + yl

plot_grid(nonlogplot, logplot, labels = "AUTO")
```



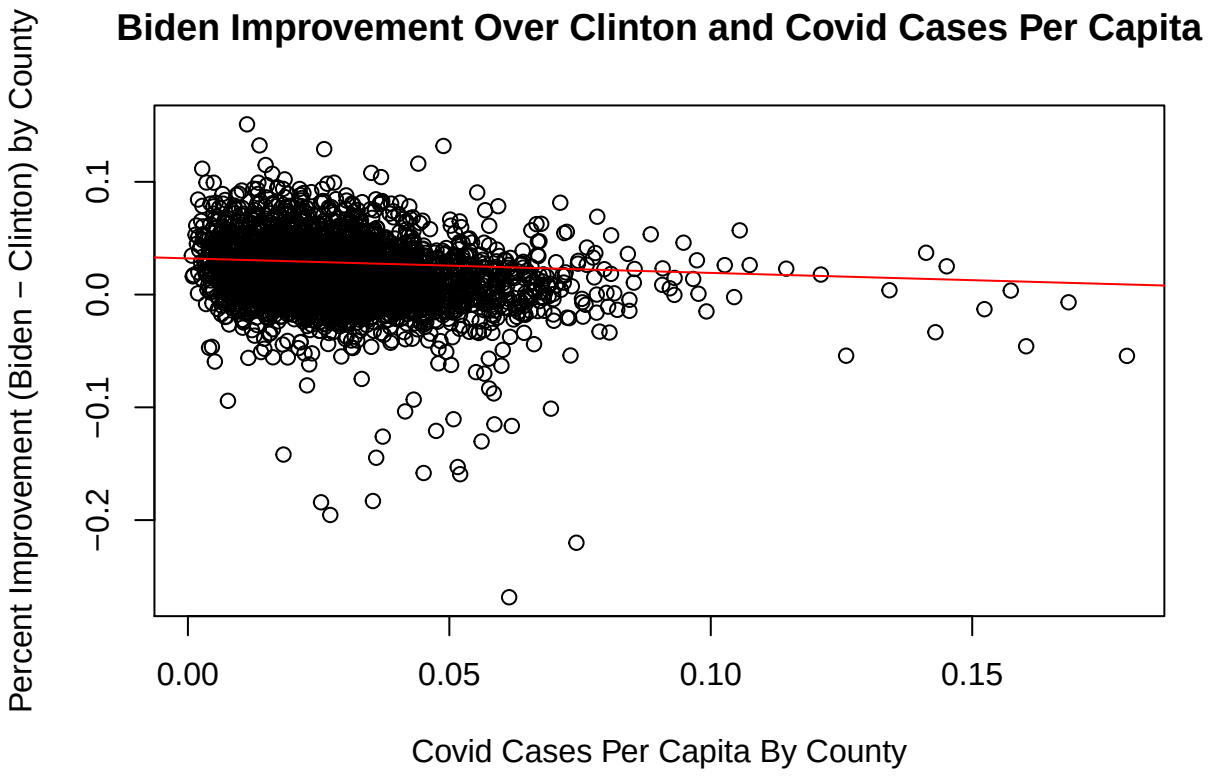
Appendix H: Can Biden's Win Be Attributed to the independent vote from 2016?

```
# find percent won in each county
both_election_data$biden_2020_percents <- both_election_data$biden_2020_votes/(both_election_data$biden
    both_election_data$trump_2020_votes + both_election_data$other_2020_votes)

both_election_data$trump_2020_percents <- both_election_data$trump_2020_votes/(both_election_data$biden
    both_election_data$trump_2020_votes + both_election_data$other_2020_votes)

both_election_data$clinton_2016_percents <- both_election_data$Votes16.Clintonh/(both_election_data$Vote
    both_election_data$Votes16.Trumpd + both_election_data$other_2016)

both_election_data$trump_2016_percents <- both_election_data$Votes16.Trumpd/(both_election_data$Votes16
    both_election_data$Votes16.Trumpd + both_election_data$other_2016)
# Initialize vectors
ttest_p_values_biden <- NULL
ttest_p_values_trump <- NULL
kstest_p_values_biden <- NULL
kstest_p_values_trump <- NULL

# add 1-100 percent of the independent vote to
# trump/clinton's 2016 totals and compare with same party
```

```r
# percentages from 2020 using KS or T test
for (i in 1:100) {
    hillary_plus_thirdparty <- (i/100) * both_election_data$other_percent_won_2016 +
        both_election_data$clinton_2016_percents
    trump_plus_thirdparty <- (i/100) * both_election_data$other_percent_won_2016 +
        both_election_data$trump_2016_percents
    ttest_p_values_biden[i] <- t.test(both_election_data$biden_2020_percents,
        hillary_plus_thirdparty)$p.value
    ttest_p_values_trump[i] <- t.test(both_election_data$trump_2020_percents,
        trump_plus_thirdparty)$p.value
    kstest_p_values_biden[i] <- ks.test(jitter(both_election_data$biden_2020_percents),
        hillary_plus_thirdparty)$p.value
    kstest_p_values_trump[i] <- ks.test(jitter(both_election_data$trump_2020_percents),
        trump_plus_thirdparty)$p.value
}
```

Appendix I: Did COVID Incidence Correlate with Biden Improving on Clinton?

```r
# Make new column for Biden Improvement
both_election_data$biden_improvement <- both_election_data$biden_2020_percents-both_election_data$clinto

# Make model for biden improvement based on covid incidence
model <- lm(cases_per_capita~biden_improvement, data = both_election_data)

# Plot
plot(both_election_data$cases_per_capita,  (both_election_data$biden_2020_percents-both_election_data$cl
     main = "Biden Improvement Over Clinton and Covid Cases Per Capita", xlab = "Covid Cases Per Capita
abline(model, col = "red")
```




```
# Display model and correlation test, both highly significant
summary(model)
```

```
##
## Call:
## lm(formula = cases_per_capita ~ biden_improvement, data = both_election_data)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.036597 -0.011822 -0.002540  0.008555  0.140478
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        0.0321049  0.0003678   87.29   <2e-16 ***
## biden_improvement -0.1289240  0.0105949  -12.17   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.01753 on 2969 degrees of freedom
## Multiple R-squared:  0.0475, Adjusted R-squared:  0.04718
## F-statistic: 148.1 on 1 and 2969 DF,  p-value: < 2.2e-16
```

```
cor.test(both_election_data$cases_per_capita, (both_election_data$biden_improvement))
```

```
##
```

```
##  Pearson's product-moment correlation
##
## data:  both_election_data$cases_per_capita and (both_election_data$biden_improvement)
## t = -12.168, df = 2969, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.2519392 -0.1834300
## sample estimates:
##        cor
## -0.2179531
```