Effect of Covid 19 on Voting Patterns in the 2020 US Presidential Election

Amir Hamilton

12/14/2020

Introduction

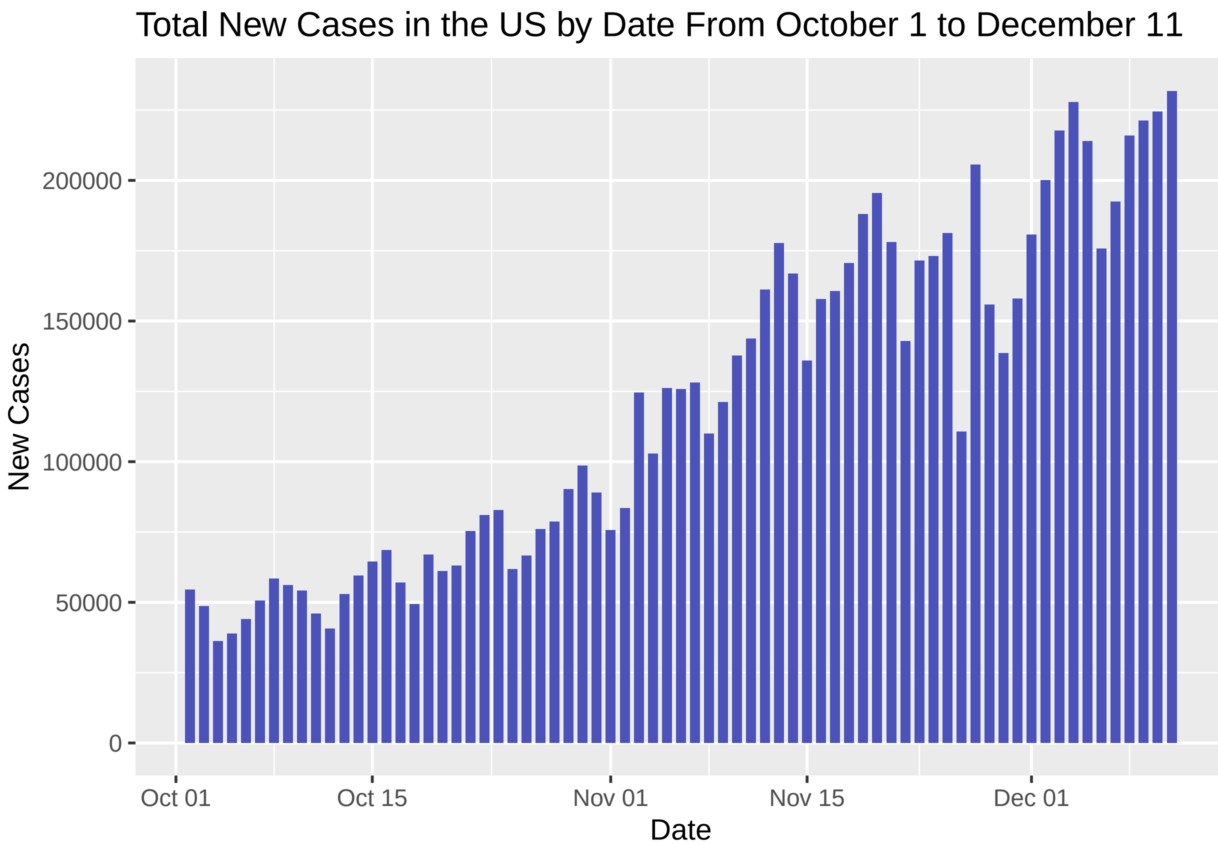
One way that social scientists are able to interact with a lay audience is by discussing the factors behind elections in the US. Some of the proposed factors include: the unequal distribution of voting machines by state, community, and race; a candidate's charisma and the perception of their character; the strength of the economy; etc. In general, these factors fall into two categories: a candidate's traits or societal trends.

This paper will look at the influence of the Covid-19 pandemic on the results of the 2020 Presidential Election. Specifically, it will examine whether there is a relationship between Covid numbers and the percentage of Trump or Biden votes on a county level.

The null hypothesis is that average treatment effect of R is 0. I expect that Covid reproduction number (R) will be a strong predictor for the percent of votes for Trump. This paper will use linear regression to estimate the treatment effect.

Covid-19 has defined the world in 2020, with millions of deaths worldwide and widespread economic chaos. In the US, there have been three main waves of infection: Spring (March and April), Summer (July), and Winter (October - Present).

To begin, a bar plot of new covid cases per day in the winter wave, which is currently ongoing, is shown on the next page. Approximately one month of data preceding election day (November 4, 2020) is shown, as are the roughly one and a half months that have passed between then and now (December 14, 2020). Data has been read in from Our World in Data's (OWID) daily updated dataset.


Total New Cases in the US by Date From October 1 to December 11
New Cases
200000
150000
100000
50000
0
Oct 01
Oct 15
Nov 01
Nov 15
Dec 01
Date

Reproduction Number (R)

The particular Covid-related statistic this paper will be looking at is R, the reproduction number.

R is a measure of the infectious potential of a disease. If R equals 1, then the number of cases will remain constant. If R is less than 1, then the number of cases are decreasing, and thus if R is greater than 1 the number of cases are increasing. Determining R is valuable for public health officials and lawmakers because it tells you what proportion of new infections you need to prevent in order to go from increasing cases to constant or decreasing cases.

At a particular time in the pandemic, Rt is written to signify the reproduction number at time t.

There are several ways to calculate R, but I'm just using the data from OWID. The formula for R is considered to be different by different epidemiological schools, and can contain numbers such as:

T – the generation time– which is the time between infection events in an infector-infectee pair of individuals;

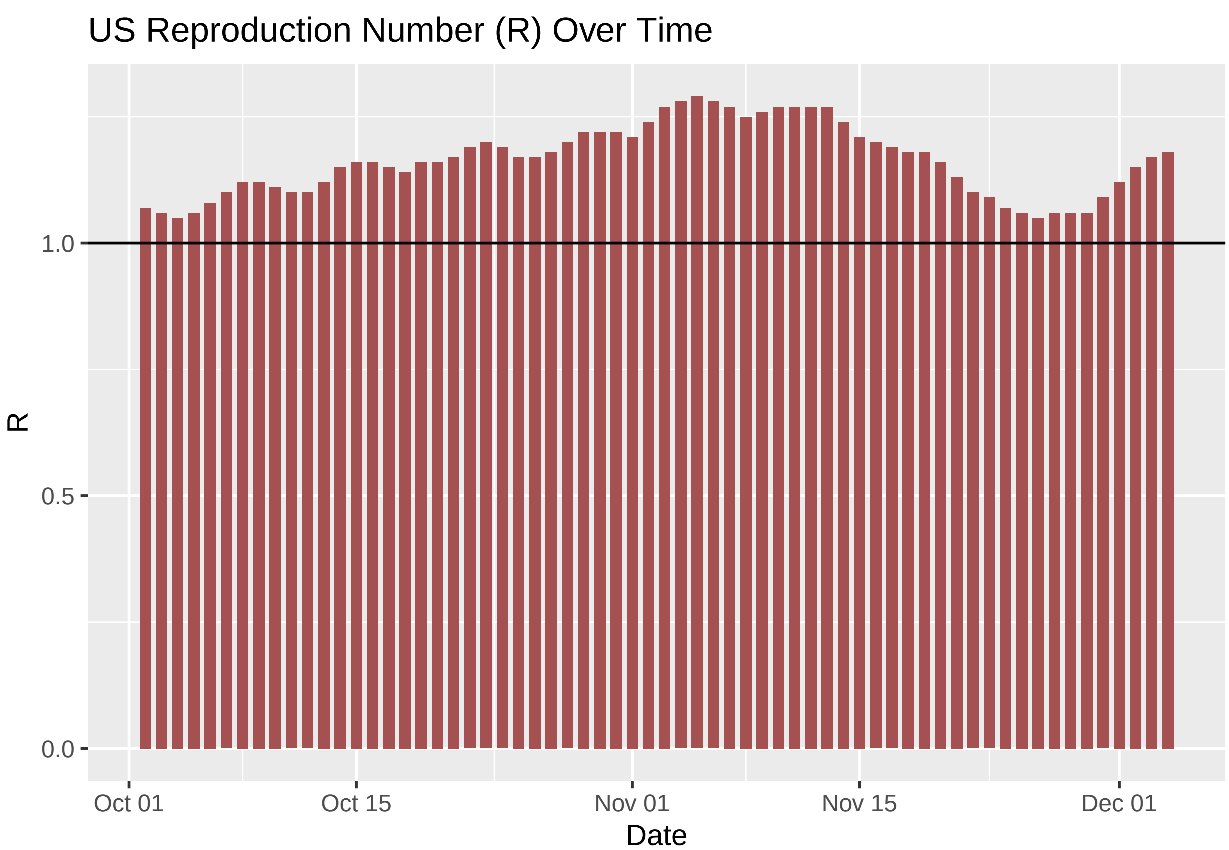
s – the serial interval – which is the average time between symptoms of infection in the transmitter to when the person they infect develops symptoms;

r – the epidemic growth rate – which is the rate at which new cases are occurring;

dt – doubling time – the amount of time which leads to a doubling of cases;

as well as the number of susceptible individuals, proportion of asymptomatic individuals, incubation time, disparate effects by age, duration from symptom onset to hospitalization or death, and the infectiousness of asymptomatic individuals. I'm leaving this calculation up to the experts, especially due to the fact that information changes frequently about these values.

Below is shown a box plot of R from the beginning of October to December 4, the latest day for which OWID has calculated R. A horizontal line of $R = 1$ shows that during the time period covered in this study, the severity of the pandemic has been increasing every day.



Presidential Election

To conduct my analysis, I had to get the data for election results. In order to achieve a sufficient level of granularity I gathered data on the county level. I could not find a dataset with all the election results for president by county, unfortunately, so I manually scraped the data from Politico and NBCNews. Another advantage of this method is that the data would be in the same format, since each state released their elections data in different formats.

Here, I read in the data from the various excel spreadsheets I created (minus Alaska and DC). Alaska was excluded because election data was unavailable on a county level and DC was excluded because there is no county there. I then merged the state/county level data to create a national dataframe of election results, which I would be able to work with.

The dataset is too large to show here in its entirety, but I've included a preview below:

```
## # A tibble: 10 x 7
##    county  trump.votes trump.percent biden.votes biden.percent state  dispID
##    <chr>         <dbl>         <dbl>       <dbl>         <dbl> <chr>  <chr>
##  1 Autauga       19838         0.716        7503         0.271 Alaba~ Autauga, ~
##  2 Baldwin       83544         0.764       24578         0.225 Alaba~ Baldwin, ~
##  3 Barbour        5622         0.535        4816         0.458 Alaba~ Barbour, ~
##  4 Bibb           7525         0.785        1986         0.207 Alaba~ Bibb, Ala~
##  5 Blount        24711         0.897        2640         0.096 Alaba~ Blount, A~
##  6 Bullock        1146         0.249        3446         0.747 Alaba~ Bullock, ~
##  7 Butler         5458         0.576        3965         0.418 Alaba~ Butler, A~
##  8 Calhoun       35101         0.69        15216         0.299 Alaba~ Calhoun, ~
##  9 Chambe~        8753         0.574        6365         0.417 Alaba~ Chambers,~
## 10 Cherok~       10583         0.861        1624         0.132 Alaba~ Cherokee,~
```

Table 1: Regression Table between Percentage of Trump Votes and R on Nov 4

	Model 1
(Intercept)	0.572***
	(0.011)
cc_nov_4$Rt_plot	0.023***
	(0.008)
Num.Obs.	1343
R2	0.006
R2 Adj.	0.005
AIC	-1105.4
BIC	-1089.8
Log.Lik.	555.685
F	8.366

* p < 0.1, ** p < 0.05, *** p < 0.01

Interactions

I decided to look at the relationship between the Percentage of Trump votes in a county or the Percentage of Biden votes in a county to the R at five different dates: Nov 4, Oct 28, Oct 21, Oct 14, and Oct 7. I chose dates that were multiples of one week prior to election day, as well as election day itself. So, I created five new dataframes created by merging the election results with the covid data on those four days.

Treatment Effect of *R* On Election Day (Nov 4) on Percentage of Trump Votes

When percentage of Trump votes per county (trump.percent) was predicted it was found that Rt_plot on election day (Nov 4) was a significant predictor (t = 2.892, p = 0.00388). The overall model fit is R^2 = 0.0062. The treatment effect is 0.023. I have shown a list of *R* values for approximately one month preceding the election.

Because the p-value is lower than the standard threshold of alpha = .05 we can reject the null hypothesis that the relationship of trump.percent to Rt_plot on Election Day is zero. This does not necessarily imply a causal relationship, however. It could be that in a county with high proportions of Trump support people are less likely to take Covid-19 seriously, or it could be that living in counties with high Rt_plot would make people more likely to vote for Trump (vote for the chaotic candidate in a chaotic circumstance).

Treatment Effect of *R* Over the Month on Percentage of Trump Votes

The data in the table shown below suggests that *R* does relate to Trump percent, with varying strengths based on the date. However, this may be due to (1) delays in reporting covid statistics and (2) a surge in people looking to get tested in the days before Thanksgiving.

Table 2: Relationship Between R and Trump Percent by Date

Date	Treatment Effect
11-04	0.023***
11-03	0.005***
11-02	0.045**
11-01	0.045**
10-31	0.023**
10-30	0.039*
10-29	3.3000000000000002E-2
10-28	3.1E-2
10-27	0.060**
10-26	0.056**
10-25	0.060***
10-24	0.054***
10-23	0.063***
10-22	2.1999999999999999E-2
10-21	-0.044**
10-20	0.033*
10-19	1.6E-2
10-18	0.059**
10-17	0.075***
10-16	0.088***
10-15	0.098***
10-14	0.088***
10-13	0.070***
10-12	0.135***
10-11	0.129***
10-10	0.091***
10-09	0.067***
10-08	0.074***
10-07	0.099***

Sources and References

Boelle, Pierre-Yves, Thomas Obadia. "Estimation of R0 and Real-Time Reproduction Number from Epidemics." 2020. https://rdrr.io/cran/R0/

Chongsuvivatwong, Virasakdi. *Analysis of epidemiological data using R and Epicalc*. 2008. Hat Yai, Thailand: Chammuang Press.

Dong, E., Du, H., & Gardner, L. (2020). An interactive web-based dashboard to track COVID-19 in real time. The Lancet infectious diseases, 20(5), 533-534.

Our World In Data. "Statistics and Research: Coronavirus Pandemic (Covid-19)." 2020. https://ourworldindata.org/coronavirus

The Royal Society. "Reproduction number (R) and growth rate (r) of the COVID-19 epidemic in the UK: methods of estimation, data sources, causes of heterogeneity, and use as a guide in policy formulation (PDF)." September 9, 2020. https://royalsociety.org/news/2020/09/set-c-covid-r-rate/

County data was manually scraped from nbcnews.com and politico.com on 2020-11-19, 2020-11-20, and 2020-12-11. Alaska was excluded due to discrepancies in the formatting of the data.

For reproduction rate by county, the dataset was too large for github so I manually reduced the size to exclude data from before November 15 and to exclude counties/dates that had too small of a sample size to produce a reproduction rate, as determined by the study.