Linkage between COVID-19 Mask Wearing and 2020 Presidential Election Votes

Term Paper

Mark Nathin & Anqi Liu

2021-12-16

Contents

Introduction

A GitHub repository where all code is hosted for this project is located at the following link:

https://github.com/marknathin/SURVMETH727_Project

The Shiny Application for this project is hosted on ShinyApps.io at the following link:

https://marknathin.shinyapps.io/SURVMETH727/

Data

Several data sources were used and combined together in order to complete this project. First, we extracted mask-wearing data from the NY Times public github repository. This was data from a survey conducted in July of 2020, around the end of the first major wave of the pandemic. The survey asked the following question: "How often do you wear a mask in public when you expect to be within six feet of another person?" The response options included Never, Rarely, Sometimes, Often, and Always. The data was compiled at the county level and includes estimated proportions of people responding to each option in each county. This data was downloaded from the NY Times Github and read in as a CSV file below.

```r
mask_use <- read_csv("~/git/SURVMETH727/mask-use-by-county.csv")
```

Next, we will read in county level results from the most recent 2020 presidential election that took place in the United States. This data was also obtained from a GitHub repository. This data was from the general election and includes both counts and proportions of voters in each county who voted democrat (for Joe Biden) or republican/gop (for Donald Trump). We will also rename "COUNTYFP" to "county_fips" to make the data joining process easier later on. This variable describes county codes that are associated with each county in the United States.

```r
Election <- read_csv("~/git/SURVMETH727/2020_US_County_Level_Presidential_Results.csv")
Election <- rename(Election, COUNTYFP = county_fips)
```

Our next data read-in consists of a CSV containing approximate latitude and longitude coordinates for all counties in the USA. We again renamed the county FIPS code variable to make it consistent with the rest of our data.

```
uscounties <- read_csv("~/git/SURVMETH727/uscounties.csv")
uscounties <- uscounties %>% rename(COUNTYFP = county_fips)
```

Our remaining variables in our dataset came from the U.S. Census Bureau and specifically the American Community Survey (ACS). The ACS is an ongoing survey that gathers a wide variety of information from households across the country. We specifically used variables from the ACS-5 which are collected every year and then averaged over a 5 year period. We used ACS-5 data collected between 2013 and 2017 which is the most recent currently available. Variables were accessed through the Census API and we used the ACS and TidyCensus packages in R to retrieve individual variables.

Before ingesting the actual data, we needed to look at the variables available. To do this, we used the `load_variables` function from the tidycensus package and specified the name of the survey (ACS-5) and year of the variables we wanted. [Alternatively, we looked through the … to locate variables]

```
acs5 <- load_variables(year = 2017, dataset = "acs5", cache = TRUE)
```

Once we view the resulting table, we could then search for key words which makes it easier to identify tables with variables of interest. For the first set of variables we used related to racial breakdowns of each county, we use the get_acs command from the tidycensus package which allows us to access variables from the census API. The geography level is specified (county level), variables are renamed to match to their variable name in the census API, and the survey from which these variables are sourced are also specified. The source here is the American Community Survey 5-year data.

```
# Race data
race <- get_acs(geography = "county",
variables = c(RaceTotal = "B02001_001", White = "B02001_002",
Black = "B02001_003", AmericanIndianAlaskaNative = "B02001_004",
Asian = "B02001_005", NativeHawaiian = "B02001_006",
OtherRace = "B02001_007", TwoRaces = "B02001_008"),
survey = "acs5")
```

```
## Getting data from the 2015-2019 5-year ACS
```

Once we have the variables we need to remove the margins of error that are linked with each variable in the census API. We then use the spread function to convert the data from a narrow to a wide format. Changing the data into this format allows us to see the data in a more intuitive way, which, in this case, puts the variable names as column headers and lists names of counties in each respective row.

```
spRace <- race %>% select(-moe) %>% spread(variable, estimate)
```

Next, each variable in this set is converted into a proportion of the county population in order to standardize it. Standardization is necessary here because some counties have a much higher population than others. Therefore, we would expect higher counts of every race in LA County compared to, say, Ziebach County in South Dakota. Some of our variables came as proportions, rather than absolute counts, but the “race” variables did not so they are standardized in the “RaceCleaned” portion of the code.

```
# Standardizing variables by county population
RaceCleaned <- spRace %>% mutate(White = White/RaceTotal,
Black = Black/RaceTotal,
AmericanIndianAlaskaNative = AmericanIndianAlaskaNative/RaceTotal,
Asian = Asian/RaceTotal,
NativeHawaiian = NativeHawaiian/RaceTotal,
TwoRaces = TwoRaces/RaceTotal) %>%
select(GEOID, NAME, RaceTotal, White, Black, AmericanIndianAlaskaNative,
Asian, NativeHawaiian, TwoRaces)
```

These same basic processes were repeated several times to obtain other variables from the census API used in this project. These include variables related to employment, median household income, and level of education. Code chunks are hidden to avoid repeat code.

Joining Data and Data Cleaning

After obtaining all of our variables, our next task was to join the data together into a single dataset that we could more easily use for analysis. We did this using the functions `inner_join` and `left_join`.

For the Census API variables, We joined by the "NAME" column each time because all of the census tables contain a columns called NAME that lists all of the county names. We joined our other data by County FIPS code.

The first chunk below joins the mask usage, election data, and county location data.

```r
Left <- left_join(Election, mask_use, by = "COUNTYFP")
Combined <- inner_join(Election, mask_use, by = "COUNTYFP")
Combined2 <- inner_join(Combined, uscounties, by = "COUNTYFP")
```

Next we do some data cleaning. This involves renaming variable names to make them more intuitive and consistent. We also create new variables including `Never_Rarely` and `Frequently_Always` which combines these two categories of the mask wearing survey into one. We did this because we thought these categories were quite similar to one another and this allows us to utilize a greater percentage of the data in our data modeling. Furthermore, we created a new variable called `Majority` which simply labels whether a higher percentage of voters in a county voted for Joe Biden or Donald Trump in the 2020 election. Finally, we round several of our variables off to two decimal places to make the numbers easier to look at in our Shiny Application.

```r
CombinedClean <- Combined2 %>%
    select(state_name.x, COUNTYFP, county_name, per_gop, per_dem, per_point_diff, NEVER,
           RARELY, SOMETIMES, FREQUENTLY, ALWAYS, lat, lng, population) %>%
  rename(Percent_Democrat = "per_dem") %>%
  rename(Percent_Republican = "per_gop") %>%
  rename(Difference = "per_point_diff") %>%
  rename(State = "state_name.x") %>%
  rename(Never = "NEVER") %>%
  rename(Rarely = "RARELY") %>%
  rename(Sometimes = "SOMETIMES") %>%
  rename(Frequently = "FREQUENTLY") %>%
  rename(Always = "ALWAYS") %>%
  mutate(Never_Rarely = Never + Rarely) %>%
  mutate(Frequently_Always = Frequently + Always) %>%
  mutate(Majority = ifelse(Percent_Democrat > Percent_Republican, "Democrat", "Republican"))

CombinedClean$Percent_Democrat <- round(CombinedClean$Percent_Democrat, digits = 2)
CombinedClean$Percent_Republican <- round(CombinedClean$Percent_Republican, digits = 2)
CombinedClean$Difference <- round(CombinedClean$Difference, digits = 2)
```

The next chunk joins all of the census data together.

```r
Join1 <- inner_join(RaceCleaned, EmploymentIncomeCleaned, by = "NAME")
Join2 <- inner_join(Join1, EducationCleaned, by = "NAME")
Join2 <- Join2 %>% select(-GEOID) %>% rename(COUNTYFP = "GEOID.x")
```

Finally, we combine our census data with all of our other data. We then rounded some additional variables to two decimal places (code hidden).

```r
CombinedClean <- left_join(CombinedClean, Join2, by = "COUNTYFP")
```

Shiny App Data

Four our Shiny Application, we selected a random sample of 1000 counties to display on our map. We selected only 1000 counties in order to avoid overcrowding the map with too many points. Furthermore, a randomly selected 1000 counties provides a fairly good representation of US counties as a whole. The plotting and data tabs on our Shiny App contains data from all counties in the USA.

Below we select the most important variables from our dataset to include in our Shiny App and save the random sample of 1000 counties as a CSV called `ShinyAppDS.csv`. The full dataset of these variables is saved as a CSV called `CombinedCleanApp.csv`.

```
ShinyAppDS <- CombinedClean %>%
  select(State, county_name, lat, lng, Percent_Republican, Percent_Democrat, Difference,
         Never, Rarely, Sometimes, Frequently, Always, Never_Rarely, Frequently_Always,
         Majority, White, Black, UnemploymentRate, Med_Household_Income, LessHS, HSGrad,
         Bachelors, GradProfDegree)

numcounties <- 1000
ShinyAppDS <- ShinyAppDS[sample(1:nrow(ShinyAppDS), numcounties, replace = FALSE),]
# Select a random sample of 1000 counties
write_csv(ShinyAppDS, "ShinyAppDS.csv")

CombinedCleanApp <- subset(CombinedClean, select = -c(19, 22, 24, 25, 28, 29, 32))
write_csv(CombinedCleanApp, "CombinedCleanApp.csv")
```

Results

This section presents the main results.

Data exploration

Now that we have compiled and cleaned all of the data we can begin exploring it. One sub-question of interest to us was which counties in the United States were most and least likely to report wearing a mask in public during the height of the pandemic? Commands from the `dplyr` package allow us to easily answer this type of question.

```
Combined %>% # Most likely to always wear mask
  select(state_name, county_name, ALWAYS) %>%
  group_by(state_name, county_name, ALWAYS) %>%
  arrange(desc(ALWAYS)) %>%
  head()
```

state_name	county_name	ALWAYS
California	Inyo County	0.889
New York	Yates County	0.884
California	Mono County	0.880
Texas	Hudspeth County	0.880
Texas	El Paso County	0.877
Nevada	Esmeralda County	0.872

We can see that Inyo County in California had an estimated 88.9% of people who would always wear a mask in public when within 6 feet of another person which is the highest in the US! Inyo County was followed closely by Yates County in New York at 88.4%. California and New York are known for being one of the most pro-democrat states in the country so this is not too surprising. However, Texas and Nevada are much more neutral on the political spectrum so it is slightly surprising to see counties in those states represented in the top 6.

```
Combined %>% # Most likely to never wear mask
  select(state_name, county_name, NEVER) %>%
  group_by(state_name, county_name, NEVER) %>%
  arrange(desc(NEVER)) %>%
  head()
```

state_name	county_name	NEVER
Utah	Millard County	0.432
Missouri	Wright County	0.419
Iowa	Cass County	0.341
Utah	Juab County	0.335
Minnesota	Jackson County	0.325
Missouri	Laclede County	0.313

On the other end of the spectrum, we can see that Millard County in Utah, Wright County in Missouri, and Cass County in Iowa led among counties for the highest proportion of people saying they would never wear a mask in public at 43.2%, 41.9%, and 34.1% respectively. Utah, Missouri, and Iowa tend to be more Republican leaning states that voted for Donald Trump in the 2020 election so this is not a huge surprise. However, these percentages are much smaller than the above percentages for counties who led the way in `always` wearing a mask. We can already see that the mask-wearing data is more skewed towards people saying that they tend to wear a mask in public when they expect to be within 6 feet of another person.

–> Look at some basic relationships

Our Shiny Application has an interactive plotting feature that allows users to explore the relationship between any two variables in our dataset.

Analysis

This section presents the main results, such as (for example) stats and graphs that show relationships, model results and/or clustering, PCA, etc.

Correlational analyses:

```
cor(CombinedClean$Never_Rarely, CombinedClean$Percent_Republican)
```

```
## [1] 0.4883538
```

```
cor(CombinedClean$Frequently_Always, CombinedClean$Percent_Democrat)
```

```
## [1] 0.5286711
```

```
cor(CombinedClean$Frequently_Always, CombinedClean$PerCapitaIncome)
```

```
## [1] 0.2720843
```

```
cor(CombinedClean$Frequently_Always, CombinedClean$Percent_Democrat)
```

```
## [1] 0.5286711
```

Linear Modeling:

Predict Percentage Vote for Democrats/Republican with Mask Wearing Tendencies

```r
## Simple Linear Regression
DemocratSupport <- lm(Percent_Democrat ~ Frequently_Always, data = CombinedClean)
summary(DemocratSupport)
```

```
##
## Call:
## lm(formula = Percent_Democrat ~ Frequently_Always, data = CombinedClean)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.40183 -0.09416 -0.01620  0.07939  0.59811
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -0.12770    0.01348  -9.474   <2e-16 ***
## Frequently_Always  0.64401    0.01854  34.733   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1357 on 3110 degrees of freedom
## Multiple R-squared:  0.2795, Adjusted R-squared:  0.2793
## F-statistic:  1206 on 1 and 3110 DF,  p-value: < 2.2e-16
```

```r
# Simple Linear Regression
RepubSupport <- lm(Percent_Republican ~ Never_Rarely, data = CombinedClean)
summary(RepubSupport)
```

```
##
## Call:
## lm(formula = Percent_Republican ~ Never_Rarely, data = CombinedClean)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.61048 -0.08325  0.01788  0.09887  0.36132
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.516672   0.004956  104.26   <2e-16 ***
## Never_Rarely 0.815080   0.026117   31.21   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1408 on 3110 degrees of freedom
## Multiple R-squared:  0.2385, Adjusted R-squared:  0.2382
## F-statistic:   974 on 1 and 3110 DF,  p-value: < 2.2e-16
```

Always wearing a mask is highly associated with a high vote percentage for democratic candidates. Never wearing a mask is associated with voting republican but the association is less strong (hypothesis: perhaps because a lot of people were wearing masks overall- perhaps fewer Republicans reported wearing a mask than Democrats but the majority of Republicans still wore masks)

```r
Mult <- lm(Percent_Democrat ~ Frequently_Always + White + Asian + Black + PerCapitaIncome + Bachelors,
           data = CombinedClean)
summary(Mult)
```

```
##
```

```
## Call:
## lm(formula = Percent_Democrat ~ Frequently_Always + White + Asian +
##     Black + PerCapitaIncome + Bachelors, data = CombinedClean)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.40020 -0.06625 -0.00200  0.06131  0.43427
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        2.221e-01  2.284e-02   9.726  < 2e-16 ***
## Frequently_Always  3.393e-01  1.486e-02  22.824  < 2e-16 ***
## White             -3.753e-01  2.226e-02 -16.860  < 2e-16 ***
## Asian              5.234e-01  8.670e-02   6.037 1.76e-09 ***
## Black              1.470e-01  2.358e-02   6.236 5.10e-10 ***
## PerCapitaIncome    1.107e-06  4.588e-07   2.412   0.0159 *
## Bachelors          9.071e-01  5.193e-02  17.469  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.09823 on 3105 degrees of freedom
## Multiple R-squared:  0.6232, Adjusted R-squared:  0.6224
## F-statistic: 855.7 on 6 and 3105 DF,  p-value: < 2.2e-16
```

Discussion

This section summarizes the results and may briefly outline advantages and limitations of the work presented. Limitation: Data collected at a single point in time but the pandemic was a rapidly changing and shifting event. Mask-mandates could impact mask-wearing survey results.

References

https://www.ers.usda.gov/data-products/county-level-data-sets/download-data

Mask-wearing data: https://github.com/nytimes/covid-19-data/tree/master/mask-use

Election data: https://github.com/tonmcg/US_County_Level_Election_Results_08-20/blob/master/2020_US_County_Level_Presidential_Results.csv