COVID19 USA

Chanho Hyun

2022 2 15

I used USA COVID19 data until 2020 Dec 10th.

I did not consider about the new variation of COVID19 such as "Omicron" so my result might be inaccurate.

Goal: Find out the relationship between the ratio of white people in each state and cases_per_100,000. (Dose white ratio affects the cases of COVID19?)

Sub Goal: If white ratio and cases_per_100,000 has no relationship, then find out the relationship between other variables.

```
df <- data.frame("variable" = c("state", "white_ratio", "cases per 100,000",
                                "ses", "sex", "oler than 65"), "Description"
                 = c("name of states", "white ratio", "cases per 100,000",
                     "socioeconomic status of the people",
" if the state has more male population than female population, 1, otherwise 0",
"ratio of people who are older than 65"))
df
```

```
##            variable
## 1             state
## 2       white_ratio
## 3 cases per 100,000
## 4               ses
## 5               sex
## 6      oler than 65
##                                                                      Description
## 1                                                                name of states
## 2                                                                   white ratio
## 3                                                             cases per 100,000
## 4                                            socioeconomic status of the people
## 5  if the state has more male population than female population, 1, otherwise 0
## 6                                         ratio of people who are older than 65
```

This is the table represents my variables. I could not get each person data who is within a state, but if there is the data about each person, then I would expect that people in the same state are not independent of each other. There is likely a region effect that would violate independence assumption. However, for this project I could not get the individual's data of each state, so we can skip this violation. (I cannot use LMM for this project)

ses is median_household_income

```r
library(readr)
library(ggplot2)
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------------------- tidyverse 1.3.0 --
```

```
## v tibble  3.0.0     v dplyr   0.8.5
## v tidyr   1.0.2     v stringr 1.4.0
## v purrr   0.3.3     v forcats 0.5.0
```

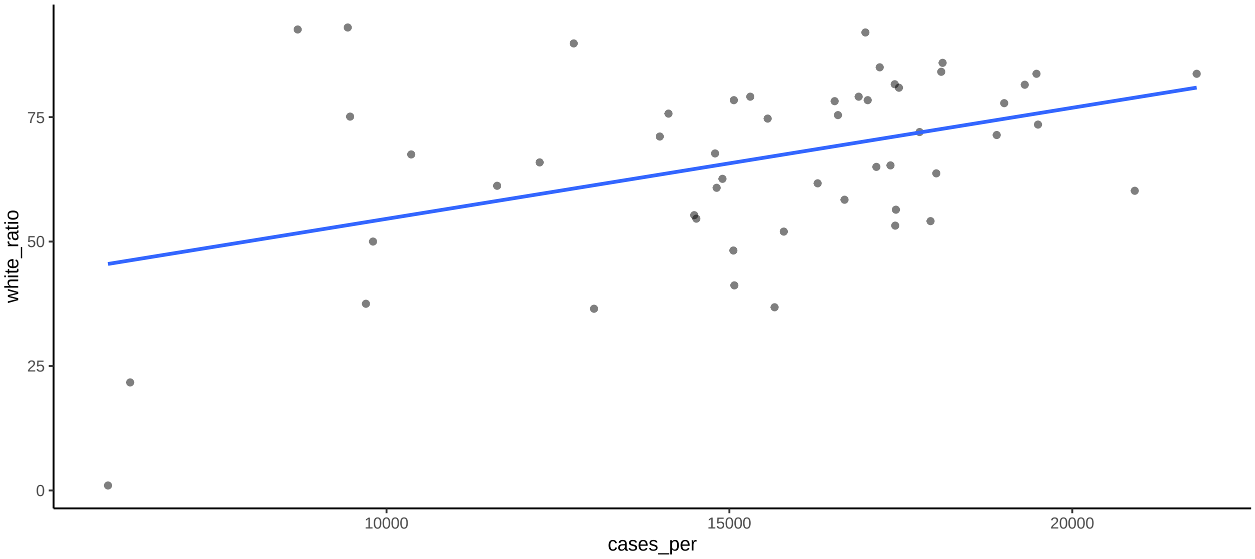
```
## -- Conflicts ------------------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```r
state <- read_csv("state.csv")
```

```
## Parsed with column specification:
## cols(
##   state = col_character(),
##   median_household_income = col_double(),
##   cases_per = col_double(),
##   older_than_65 = col_double(),
##   sex = col_double(),
##   white_ratio = col_double()
## )
```

```r
ggplot(data=state, aes(x=cases_per, y=white_ratio)) +
  geom_point(alpha = 0.5) + geom_smooth(method = "lm", se = FALSE) +
  theme_classic()
```

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



This a scatter plot to examine the relationship between casese per 100,000 and white_ratio Include a line of best fit.

Followed by scatter plot, line of best fit shows there is a reasonable relationship between cases_per_100,000 and minotirty_status since it has positive slope. However, if there is no line of best fit, it is hard to find out the relationship between white_ratio and cases_per because dots are too scattered.

```
lm <- lm(white_ratio ~ cases_per + sex + median_household_income +
           older_than_65, data = state)

summary(lm)
```

```
##
## Call:
## lm(formula = white_ratio ~ cases_per + sex + median_household_income +
##     older_than_65, data = state)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -32.849  -9.610   1.259  11.350  35.931
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)             -7.308e+01  3.717e+01  -1.966   0.0552 .
## cases_per                3.138e-03  7.150e-04   4.389 6.41e-05 ***
## sex                      3.902e-01  5.567e+00   0.070   0.9444
## median_household_income  4.588e-04  2.215e-04   2.071   0.0438 *
## older_than_65            3.700e+00  1.296e+00   2.854   0.0064 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.99 on 47 degrees of freedom
## Multiple R-squared:  0.3115, Adjusted R-squared:  0.2529
## F-statistic: 5.316 on 4 and 47 DF,  p-value: 0.00129
```

This is linear model with test as the response and use cases_per_100,000, sex, median_household_income, ratio of people who are older than 65 as the covariates.Followed by p-value, cases_per, median_household_income, and older_than_65 are statistically significant, but sex is not.

```
confint(lm)
```

```
##                                 2.5 %       97.5 %
## (Intercept)             -1.478628e+02 1.702814e+00
## cases_per                1.699815e-03 4.576623e-03
## sex                     -1.080879e+01 1.158923e+01
## median_household_income  1.319329e-05 9.044082e-04
## older_than_65            1.091793e+00 6.307976e+00
```

Followed by confint(lm), my (intercept) which is 95% confidence interval of white ratio has negative lower bound. So, I put zero instead of negative lower- bound. Thus, we can claim that with 95% conficence, our intercept is between 0 and 1.7028e+00.

cases_per looks statistically significant. With 95% of confidence, one unit increase in cases_per, white_ratio is changed from 1.6998e-03 to 4.5766e-03

As a result, we can confirm that cases_per and white_ratio do have relationship. However, we got negative lower bound for our 95% confidence of intercept and the differences between cases_per's lower bound and upper bound is too small. Therefore, we cannot claim that the ratio of white people affects cases_per.

I wanted to use linear mixed model function for my project, but I could not get the data from each person in each state, so I could not use lmm for this project.

citation:

Covid-19 in the United States. Data USA. (n.d.). Retrieved February 15, 2022, from https://datausa.io/coronavirus

U.S. Census Bureau quickfacts: California. (n.d.). Retrieved February 15, 2022, from https://www.census.gov/quickfacts/fact/table/CA/PST045221