Covid-19 death and vaccination base on time series data

Yao Lu

3/17/23

# 1. Summary/Abstract

*Write a summary of your project.*

Do regression for the relationship between the covid infection and vaccination. Maybe add some other variables like population, age, gender, race. Plan to use county as individual observation.

# 2. Introduction

## 2.1 General Background Information

*The main data we use is the vaccination data and the data of the COVID cases in Georgia. Then after we make a summary of the data. We try some models which could guide us to do further study.Our research question is, does the coverage of vaccine decrease the spread of COVID. Under the background of the pandemic, people generate the vaccine to prevent the spread of the COVID. How many effect do the vaccine have? Now, the method we use here is using crude model to briefly test the relationship between coverage of the vaccine and incerasing speed of the COVID.*

## 2.2 Description of data and data source

*Firstly, we load data and see if there is the need of clean and transformation*

All data wrangling process are recorded in code/processing\_code.

covid vaccination data comes from https://experience.arcgis.com/experience/3d8eea39f5c1443db1743a4cb8948a9c covid symptom and death data comes from https://ga-covid19.ondemand.sas.com/docs/ga\_covid\_data.zip

All the raw data are saved at raw\_data

## 2.3 Questions/Hypotheses to be addressed

Can booster decrease the speed of COVID spread?

# 3. Methods

Start from linear regression. Then may try CV/bootstraping.

## 3.1 Data import and cleaning

## 3.2 Statistical analysis

# 4. Results

## 4.1 Exploratory/Descriptive analysis

Load the packages

#load packages  
library(here)  
library(dplyr)  
library(ggplot2)

Load the data.

#Path to data. Note the use of the here() package and not absolute paths  
data\_location <- here::here("data","processed\_data","processeddata.rds")  
#load data  
mydata <- readRDS(data\_location)

### 4.1.1 Data exploration through tables

Showing a bit of code to produce and save a summary table.

summary\_df = skimr::skim(mydata[,c('14 day case rate','RTCUMVAXADMIN',  
 'PCTCUMPVAX',"PCTCUMPCVAX","PCTBOOSTER")])  
print(summary\_df)

── Data Summary ────────────────────────  
 Values  
Name ...[]   
Number of rows 159   
Number of columns 5   
\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_   
Column type frequency:   
 numeric 5   
\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_   
Group variables None   
  
── Variable type: numeric ──────────────────────────────────────────────────────  
 skim\_variable n\_missing complete\_rate mean sd p0 p25  
1 14 day case rate 0 1 83.4 97.3 0 41.8  
2 RTCUMVAXADMIN 0 1 127918. 27260. 60435 109456.   
3 PCTCUMPVAX 0 1 51.4 9.48 29.6 45.4  
4 PCTCUMPCVAX 0 1 46.9 8.88 24.9 41.5  
5 PCTBOOSTER 0 1 20.8 5.73 7 17   
 p50 p75 p100 hist   
1 74 104. 1126. ▇▁▁▁▁  
2 127981 143319 233717 ▂▇▇▁▁  
3 51.5 56.7 92.5 ▂▇▅▁▁  
4 47 52.1 83.1 ▂▇▆▁▁  
5 21 24 39 ▂▆▇▁▁

# save to file  
summarytable\_file = here("results","exploratory", "summarytable.rds")  
saveRDS(summary\_df, file = summarytable\_file)

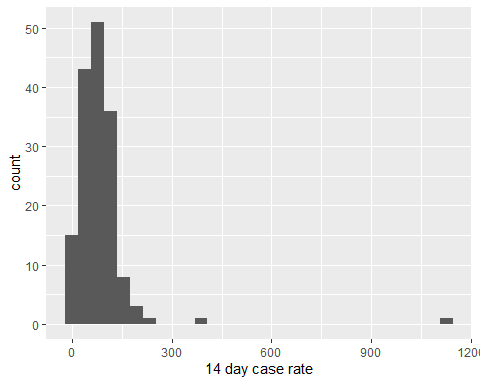
### 4.1.2 Data exploration through figures

Histogram plots for the continuous outcomes.

‘14 day case rate’ first.

p1 <- mydata %>% ggplot(aes(x=`14 day case rate`)) + geom\_histogram()   
plot(p1)

`stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



figure\_file = here("results","exploratory","14\_day\_case\_rate.png")  
ggsave(filename = figure\_file, plot=p1)

Saving 5 x 4 in image  
`stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

Here we find there are two county which have high ‘14 day case rate’. We want to know which they are.

mydata$county\_name[which(mydata$`14 day case rate`>300)]

[1] "Quitman" "Stewart"

mydata$`14 day case rate`[which(mydata$county\_name=='Quitman')]

[1] 392.33

mydata$`14 day case rate`[which(mydata$county\_name=='Stewart')]

[1] 1125.8

They are ‘Quitman’ and ‘Stewart’ counties. In the previous 14 days/100,000 population, Quitman has 392.33 cases reported, and Stewart has 1125.8 cases reported. Beyond our current research, we should pay attention about what situations make these two counties have such a high level than other counties. Previous 14 days before 01/31/2023.

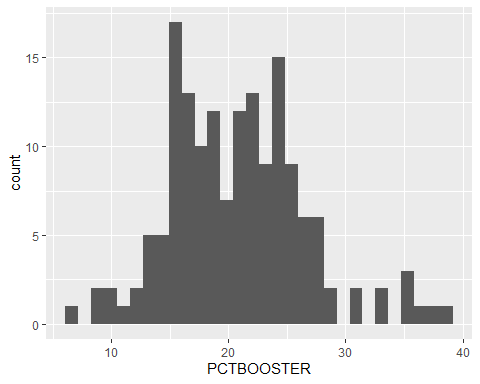
Now population, RTCUMVAXADMIN,PCTCUMPVAX,PCTCUMPCVAX and PCTBOOSTER. Since last four should be highly correlated. The analysis we will apply later should be same. Our final model should include two predictors which are population and one of these four.

Since the four choices share same statistics steps. We just pick population+PCTBOOSTER as example. I was thinking where could I find a categorical variables. Such as could I make population as ‘small’,‘medium’ and ‘large’. I think we can try both way how we define population. Since there is not other good choice of categorical variable I can find in this dataset. To make the diversity of our predictors. I decide to define population as categorical variables.(However,I am not saying this a better way.)

Now ‘PCTBOOSTER’.

px5 <- mydata %>% ggplot(aes(x=PCTBOOSTER)) + geom\_histogram()   
plot(px5)

`stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



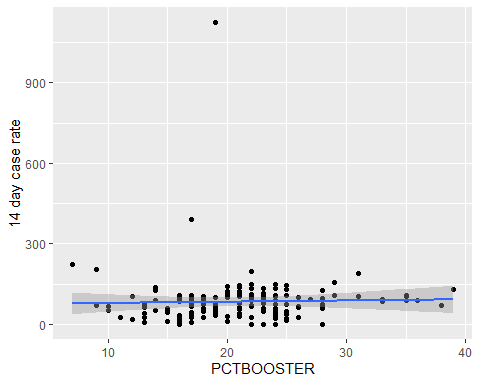
figure\_file = here("results","exploratory","PCTBOOSTER\_distribution.png")  
ggsave(filename = figure\_file, plot=px5)

Saving 5 x 4 in image  
`stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

Now ‘14 day case rate’ as function of ‘PCTBOOSTER’.

pf1 <- mydata %>% ggplot(aes(x=PCTBOOSTER, y=`14 day case rate`)) + geom\_point() + geom\_smooth(method='lm')  
plot(pf1)

`geom\_smooth()` using formula 'y ~ x'



figure\_file = here("results","exploratory","case\_BOOSTER.png")  
ggsave(filename = figure\_file, plot=pf1)

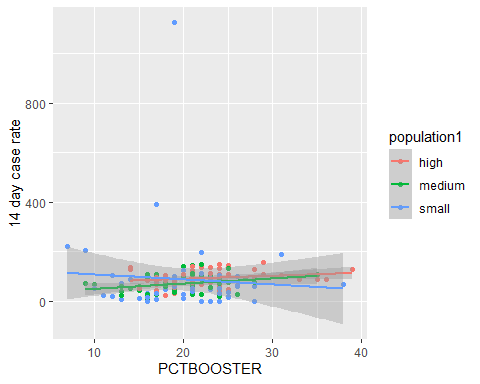
Saving 5 x 4 in image  
`geom\_smooth()` using formula 'y ~ x'

Here we find that two outliers make our plot hard to see clearly. We can try another model removing these two outliers. Here I just leave it here to see what our model shows.

Once more cases as function of PCTBOOSTER, stratified by population.

pf2 <- mydata %>% ggplot(aes(x=PCTBOOSTER, y=`14 day case rate`, color = population1)) + geom\_point() + geom\_smooth(method='lm')  
plot(pf2)

`geom\_smooth()` using formula 'y ~ x'



figure\_file = here("results","exploratory","case\_BOOSTER\_stratified.png")  
ggsave(filename = figure\_file, plot=pf2)

Saving 5 x 4 in image  
`geom\_smooth()` using formula 'y ~ x'

### 4.1.3 Notes

Here we can see both the two outliers have small population.

## 4.2 Basic statistical analysis

Load the packages

#load packages  
library(here)  
library(dplyr)  
library(ggplot2)

Load the data.

#Path to data. Note the use of the here() package and not absolute paths  
data\_location <- here::here("data","processed\_data","processeddata.rds")  
#load data  
mydata <- readRDS(data\_location)

Linear regression

# linear regression  
lm1 <- lm(`14 day case rate`~PCTBOOSTER+population1,data=mydata)  
  
summary(lm1)

Call:  
lm(formula = `14 day case rate` ~ PCTBOOSTER + population1, data = mydata)  
  
Residuals:  
 Min 1Q Median 3Q Max   
 -88.77 -38.79 -7.89 19.12 1038.57   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) 88.9698 34.4207 2.585 0.0107 \*  
PCTBOOSTER 0.1713 1.3962 0.123 0.9025   
population1medium -22.8006 19.4150 -1.174 0.2420   
population1small -4.9922 19.3080 -0.259 0.7963   
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 97.68 on 155 degrees of freedom  
Multiple R-squared: 0.01056, Adjusted R-squared: -0.008594   
F-statistic: 0.5512 on 3 and 155 DF, p-value: 0.6481

lm2 <- lm(`14 day case rate`~PCTCUMPCVAX+population1,data=mydata)  
  
summary(lm2)

Call:  
lm(formula = `14 day case rate` ~ PCTCUMPCVAX + population1,   
 data = mydata)  
  
Residuals:  
 Min 1Q Median 3Q Max   
 -92.30 -38.83 -11.39 21.39 1032.29   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) 126.5108 47.1870 2.681 0.00813 \*\*  
PCTCUMPCVAX -0.6718 0.9041 -0.743 0.45859   
population1medium -26.7886 19.5176 -1.373 0.17188   
population1small -8.4110 19.2612 -0.437 0.66295   
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 97.51 on 155 degrees of freedom  
Multiple R-squared: 0.01397, Adjusted R-squared: -0.005112   
F-statistic: 0.7321 on 3 and 155 DF, p-value: 0.5343

## 4.3 Full analysis

# 5. Discussion

## 5.1 Summary and Interpretation

## 5.2 Strengths and Limitations

## 5.3 Conclusions

# 6. References