Covid-19 death and vaccination base on time series data

Yao Lu

2/24/23

The structure below is one possible setup for a data analysis project (including the course project). For a manuscript, adjust as needed. You don’t need to have exactly these sections, but the content covering those sections should be addressed.

This uses MS Word as output format. [See here](https://quarto.org/docs/output-formats/ms-word.html) for more information. You can switch to other formats, like html or pdf. See [the Quarto documentation](https://quarto.org/) for other formats.

# 1. Summary/Abstract

*Write a summary of your project.*

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

Here I plan to explore these documents.

Here is the initial plan. First do ARIMA model to check the best time series model for the infection(may measured by death or symptom, haven’t decided)

Second, 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*

library(readxl)  
library(readr)  
library(DataExplorer)  
library(dplyr)

Attaching package: 'dplyr'

The following objects are masked from 'package:stats':  
  
 filter, lag

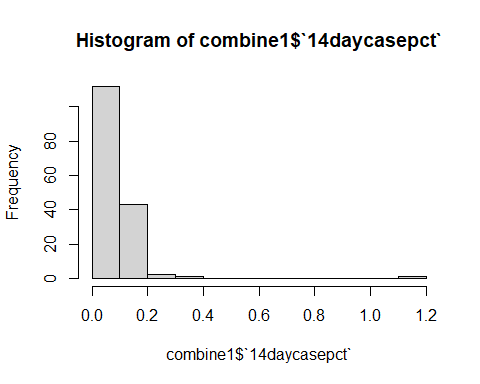
The following objects are masked from 'package:base':  
  
 intersect, setdiff, setequal, union

library(ggplot2)  
library(here)  
  
data\_location <- here::here("data","raw\_data","Georgia\_DPH\_PUBLIC\_Vaccination\_Public\_Data\_in\_Excel.xlsx")  
vacin <- readxl::read\_excel(data\_location,sheet=2)  
  
data\_location1 <- here::here("data","raw\_data",'ga\_covid\_data',"county\_cases.csv")  
  
cases <- read\_csv(data\_location1)

Rows: 161 Columns: 15

── Column specification ────────────────────────────────────────────────────────  
Delimiter: ","  
chr (2): county\_name, county\_id  
dbl (13): cases, State FIPS code, County FIPS code, population, confirmed\_ca...  
  
ℹ Use `spec()` to retrieve the full column specification for this data.  
ℹ Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

#Both vacin and cases have some unknown,so we drop them  
vacin <- vacin %>% filter(COUNTY\_ID!=99999)  
vacin <- vacin %>% filter(COUNTY\_NAME!="GEORGIA")  
cases <- cases %>% filter(population!=0)  
  
temp <- vacin[44,]  
vacin[44,] <- vacin[43,]  
vacin[43,] <- temp  
  
combine <- cbind(vacin,cases)  
  
#There are many variables which we won't need. So we clean some of them  
combine1 <- combine %>% dplyr::select(COUNTY\_NAME,PERSONVAX,PCTCUMPVAX, population, `14 day case rate`, `14 day cases`)  
#Prepare for the two plot, one is absolute plot(None vac person vs 14 absolute number)  
#the other is percentage plot (None vac percentage vs 14 day cases percentage)  
combine1$novac <- combine1$population-combine1$PERSONVAX  
combine1$PCTCUMPVAX <- as.numeric(combine1$PCTCUMPVAX)  
  
  
complement <- function (x) {  
 x <- 100-x  
}  
  
combine1$novacpct <- complement(combine1$PCTCUMPVAX)  
  
#this are subsituted by complement function  
#for(i in 1:nrow(combine1)) {  
# combine1$novacpct[i] <- 100-as.numeric(combine1$PCTCUMPVAX[i])  
#}  
  
per1000 <- function(x){  
 x <- x/1000  
}  
  
combine1$`14daycasepct` <- per1000(combine1$`14 day case rate`)  
  
#for(i in 1:nrow(combine1)) {  
# combine1$`14daycasepct` <- combine1$`14 day case rate`[i]/1000  
#}  
  
#now we get the data we want  
#View(combine1)  
  
hist(combine1$`14daycasepct` )



## 2.3 Questions/Hypotheses to be addressed

# 3. Methods

## 3.1 Data aquisition

## 3.2 Data import and cleaning

## 3.3 Statistical analysis

# 4. Results

## 4.1 Exploratory/Descriptive analysis

## 4.2 Basic statistical analysis

## 4.3 Full analysis

# 5. Discussion

## 5.1 Summary and Interpretation

## 5.2 Strengths and Limitations

## 5.3 Conclusions

# 6. References