Final

Qing Dai Hongyi Duan Yili Luo Tiancheng Pan Jinshui Zhang Shusheng Zhang

2022-04-18

Introduction

With continuously changing abilities and accumulating mutations, SARS-CoV-2, the virus that causes COVID-19, constant evolvements and accumulated mutations in its genetic code over time. The emergence and quick spread of the alpha, beta, and delta SARS-CoV-2 VOCs have generated continuous waves of infection in the past two years. The virus has brought tremendous shocks to the supply side of the economy and resulted in millions of deaths around the globe, representing an unprecedented tragic loss of the whole human society.

By analyzing the Covid-19 Case Surveillance Public Use Data from the Centers for Disease Control and Prevention, our project aims to identify the primary factors that are sensible to the effects of Covid-19. We mainly focus on samples in North Carolina and ignore the individual observations which have missing/unknown live status records. Focusing on a single state would help eliminate potential time-invariant effects among different states.

We hope this project will bring suggestive policy implications by identifying the most vulnerable groups against the Covid-19 virus among the population. Hopefully with our convincing results, the medical facilities would be able to allocate resources, such as hospitalization and medical aids, to the appropriate groups efficiently. Also, the government can assign social welfare benefits and designate priorities for vaccination by understanding which group is most vulnerable to the virus.

FAQ

Where is the data come from and how to clean it?

The Data is come from CDC COVID-19-Case-Surveillance-Public-Use-Data Form, which can be found at https://data.cdc.gov/Case-Surveillance/COVID-19-Case-Surveillance-Public-Use-Data-with-Ge/n8mc-b4w4/data. The code table and introduction for each parameter can also be found at the website. After filtered all the NC 2020 observation with None-missing value of `death_yn`, there are totally 575400 observation with 19 variables, and the summary of the data set are in the appendix.

Data cleaning procedure:

- Drop `res_state`, `state_fips_code` since all the observations come from NC
- Drop `county_fips_code` since it is redundant as `res_county`
- Drop `case_onset_interval`, `process`, `exposure_yn`, `-icu_yn`, `underlying_conditions_yn` since they have highly rate of missing or unknown
- Drop the observation with `NA`
- Combine the `Missing` and `Unknown` of `symptom_status` as `unknown`
- Convert the `death_yn` as binary variable with `Yes` as 1
- Change all the categorical data as factor

After all we have 425307 observations left with 11 variables. The summary of the new data set can also be found at appendix

Reference

Bang, H. and Robins, J. (2005). Doubly Robust Estimation in Missing Data and Causal Inference Models, Biometrics, Volume 61, Pages 962–972. DOI: 10.1111/j.1541-0420.2005.00377.x

CDC. COVID 19 Case Surveillance Public Use Data Form. https://data.cdc.gov/Case-Surveillance/COVID-19-Case-Surveillance-Public-Use-Data-with-Ge/n8mc-b4w4/data

Appendix

Summary of Original Data

```
##    case_month     res_state   state_fips_code       res_county
##  2020-12:195568   NC:575400   Min.   :37      MECKLENBURG: 66644
##  2020-11:103453               1st Qu.:37      WAKE       : 45793
##  2020-10: 63455               Median :37      GUILFORD   : 26375
##  2020-07: 54015               Mean   :37      FORSYTH    : 21623
##  2020-08: 43620               3rd Qu.:37      GASTON     : 16321
##  2020-06: 42996               Max.   :37      (Other)    :382586
##  (Other): 72293                               NA's       : 16058
##  county_fips_code          age_group           sex
##  Min.   :37001    0 - 17 years  : 64882   Female :292579
##  1st Qu.:37063    18 to 49 years:311481   Male   :254172
##  Median :37109    50 to 64 years:114004   Missing:   563
##  Mean   :37105    65+ years     : 76362   NA's   : 28086
##  3rd Qu.:37151    NA's          :  8671
##  Max.   :37197
##  NA's   :16058
##                             race                      ethnicity
##  White                        :282065   Hispanic/Latino    : 38215
##  Black                        : 81269   Non-Hispanic/Latino:257526
##  Unknown                      : 79637   Unknown            :140826
##  Asian                        :  4735   NA's               :138833
##  American Indian/Alaska Native:  4110
##  (Other)                      :  1884
##  NA's                         :121700
##  case_positive_specimen_interval case_onset_interval
##  Min.   :-40.000                 Min.   :-16
##  1st Qu.:  0.000                 1st Qu.:  0
##  Median :  0.000                 Median :  0
##  Mean   :  0.193                 Mean   :  0
##  3rd Qu.:  0.000                 3rd Qu.:  0
##  Max.   :105.000                 Max.   : 22
##  NA's   :195                     NA's   :298698
##                            process        exposure_yn
##  Clinical evaluation            :    46   Missing:264105
##  Contact tracing of case patient:     8   Unknown:259597
##  Missing                        :575334   Yes    : 51698
##  Multiple                       :     6
##  Other                          :     6
```

```
##
##
##                    current_status        symptom_status      hosp_yn
##  Laboratory-confirmed case:520209   Asymptomatic: 58823   Missing:    52
##  Probable Case            : 55191   Missing     :   117   No     :354159
##                                     Symptomatic :333915   Unknown:200548
##                                     Unknown     :182545   Yes    : 20641
##
##
##
##      icu_yn       death_yn     underlying_conditions_yn
##  Missing:   235   No :573258      :476449
##  No     :  6291   Yes:  2142   Yes: 98951
##  Unknown:566279
##  Yes    :  2595
##
##
##
```

Summary of the cleaning Data

```
##    case_month           res_county              age_group           sex
##  2020-12:160678   MECKLENBURG: 56934   0 - 17 years  : 38856   Female :228729
##  2020-11: 81076   WAKE       : 35748   18 to 49 years:244613   Male   :196313
##  2020-10: 46758   GUILFORD   : 22769   50 to 64 years: 83775   Missing:   265
##  2020-07: 37481   FORSYTH    : 17461   65+ years     : 58063
##  2020-08: 28771   GASTON     : 14049
##  2020-06: 28206   CUMBERLAND : 12073
##  (Other): 42337   (Other)    :266273
##                             race                     ethnicity
##  American Indian/Alaska Native:  3523   Hispanic/Latino    : 38005
##  Asian                         :  4228   Non-Hispanic/Latino:249948
##  Black                         : 74821   Unknown            :137354
##  Multiple/Other                :  1768
##  Unknown                       : 75773
##  White                         :265194
##
##  case_positive_specimen_interval                   current_status
##  Min.   :-29.0000                Laboratory-confirmed case:380472
##  1st Qu.:  0.0000                Probable Case            : 44835
##  Median :  0.0000
##  Mean   :  0.1693
##  3rd Qu.:  0.0000
##  Max.   : 89.0000
##
##      symptom_status      hosp_yn          death_yn
##  Asymptomatic: 40990   No     :257828   Min.   :0.000000
##  Symptomatic :242555   Unknown:153562   1st Qu.:0.000000
##  Unknown     :141762   Yes    : 13917   Median :0.000000
##                                         Mean   :0.004745
##                                         3rd Qu.:0.000000
##                                         Max.   :1.000000
##
```

code

```
knitr::opts_chunk$set(echo = TRUE)
library(tidyverse)
library(ggplot2)
library(dplyr)
library(zoo)
library(readxl)
library(cobalt)
library(sandwich)
library(PSweight)
library(knitr)
library(lme4)
library(boot)
library(stan4bart)
library(rstan)
library(kableExtra)
covid=read.csv("COVID-19_Case_Surveillance_Public_Use_Data_with_Geography.csv",header = TRUE)
factor_covid <- as.data.frame(unclass(covid),stringsAsFactors=TRUE)
covid_clean=covid%>%
  select(-res_state,-state_fips_code,-county_fips_code,-case_onset_interval,-process,-exposure_yn, -icu
  drop_na()%>%
  mutate(symptom_status=ifelse(symptom_status=="Missing","Unknown",symptom_status),
         death_yn=ifelse(death_yn=="Yes",1,0))
covid_clean=as.data.frame(unclass(covid_clean),stringsAsFactors=TRUE)
summary(factor_covid)
summary(covid_clean)
```