

# 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, have constant evolvments 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 whether the hospitalized are sensible to the effects of Covid-19 under Double Robust Estimator. 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

### Q1 What is Double Robust estimator and why use it?

Double Robust estimator is a estimator of average treatment effect under causal inference:

$$\begin{aligned}\tau^{DR} = & N^{-1} \sum_{i=1}^N \left\{ \hat{m}_1(X_i) + \frac{Z_i \{Y_i - \hat{m}_1(X_i)\}}{\hat{e}(X_i)} \right\} \\ & - N^{-1} \sum_{i=1}^N \left\{ \hat{m}_0(X_i) + \frac{(1 - Z_i) \{Y_i - \hat{m}_0(X_i)\}}{1 - \hat{e}(X_i)} \right\}\end{aligned}$$

Where:

- $\hat{e}(X_i)$  is the estimated Propensity score for each observation
- $\hat{m}_j(X_i)$  for  $j \in \{0, 1\}$  is the estimated outcome model for each observation in either treatment or control group
- $Y_i$  is the outcome variable `death_yn`
- $Z_i$  is the treatment variable `hosp_yn`

- $X_i$  is the covariates, including sex, age, and etc.

Compared to just focusing on estimators of any Machine Learning model, we choose the causal inference since we want to make sure the effect of hospitalization on the death rate is on itself instead of on any other confounders. Moreover, according to theoretical proofs and simulations from Bang and Robins (2005), the double robust estimator provides two opportunities to deduce nearly correct inference on the causal effect. This method, in turn, indicates that even if one of the outcome models or Propensity score models is wrong, we can still accurately obtain an unbiased causal effect.

## Q2 Where does the data come from?

Our data comes from CDC COVID-19-Case-Surveillance-Public-Use-Data Form, which can be accessed at the website below: <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 on the website. After filtering all the NC 2020 observations with non-missing values of `death_yn`, we identify that there are 575400 observations with 19 variables. The summary of the data set is in the appendix. Specifically,

- `res_county`: The county of the observation
- `case_month`: The Date received by CDC
- `age_group`: This is a categorical variable which has three values: 0-17 years; 18-49 years; 50-64 years; 65+years
- `race`: This is another categorical variable which has six values: American Indian/Alaska Native; Asian; Black; Multiple/Other; Native Hawaiian/Other Pacific Islander; White
- `sex`: This is a variable which has three values: Male; Female; Other
- `ethnicity`: This is a variable which has two values: Hispanic/Non-Hispanic
- `hosp_yn`: Was this patient hospitalized? Yes/No/Unknown
- `death_yn`: Did the patient die as a result of this illness? Yes/No

## Q3 Why you choose this Data source than others?

We choose this data source because compared to others, it includes all cases with the earliest date available in each record (date received by CDC or date related to illness/specimen collection) at least 14 days before constructing the previously updated datasets. This 14-day lag stabilizes case reporting and ensures that time-dependent outcome data is accurately captured.

Moreover, most other data sources only provide the death cases without the survive status, which is hard for us to build models just on the death rate.

## Q4 Why choose North Carolina? Why not choose the United States?

We choose the state of North Carolina to eliminate potential bias existing in the large volume of data. When including every state in the country, the data contains 1.8 billion observations during our time interval. If so, we will have no methods to reduce the time-invariant effect lying in each state, such as geography variance, population structure differences, and government efficiencies data has no measures. Moreover, to include all the states, we may consider both the fixed and random group effects caused by different states, which may cause our model too complicated to converge. Therefore, we choose to explore North Carolina.

## Q5 Why 2020?

We focus on the year 2020 to eliminate the potential effects of vaccination. According to the reports from CDC, North Carolina's COVID-19 Vaccine eligibility opens for all adults on April 7. Since our data source does not contain information regarding the individual's vaccination status, we only focus on the year 2020.

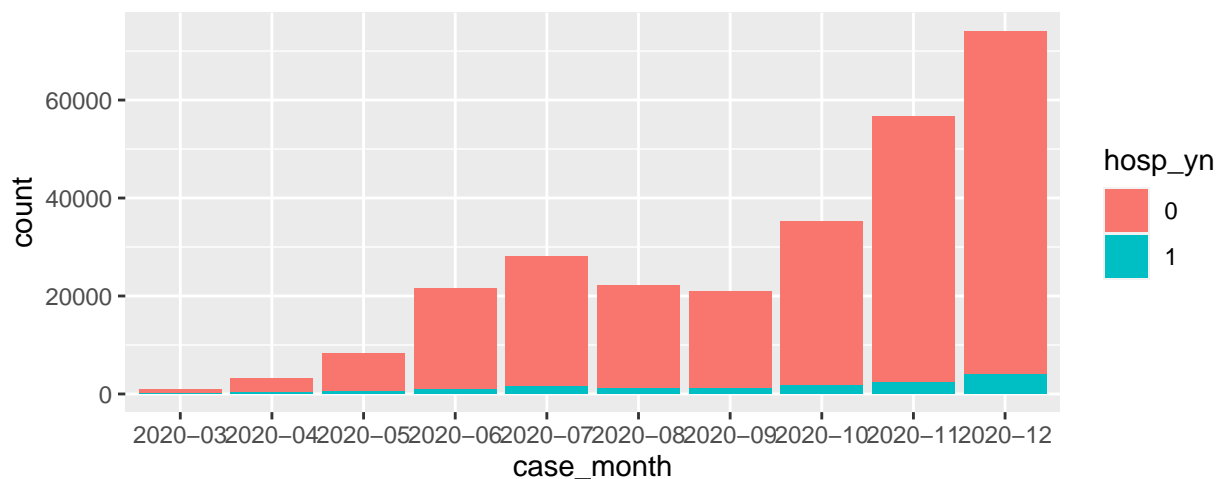
## Q6 Any Data cleaning before modeling?

Data cleaning procedure:

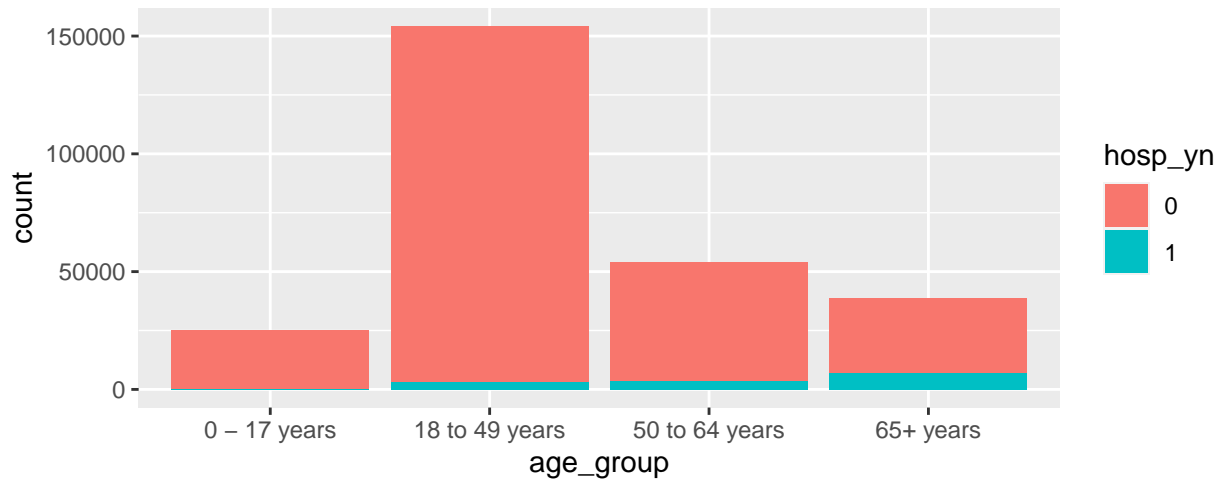
- Drop `res_state`, `state_fips_code` since all the observations come from NC
- Drop `county_fips_code` since it is redundant as `res_county`
- Drop `case_onset_interval`, `process`, `exposure_yn`, `-icu_yn`, `underlying_conditions_yn` since they have highly rate of missing or unknown
- Drop the observation with NA
- Combine the Missing and Unknown of `symptom_status` as unknown
- Drop the Unknown of `hosp_yn`
- Convert the `death_yn`, and `hosp_yn` as binary variable with Yes as 1
- Drop the `case_positive_specimen_interval`, `current_status`, `symptom_status` due to the complexity of the model
- Drop the Missing in `Sex` due to the limited number of observations
- Combine the county with county with 1
- Change all the categorical data as factor
- Combine the County with 2000 or less observation as **other** to reduce the group effect

After cleaning the first part, we get 271822 observations with 270088 survivors and 1734 deaths which can be found in the appendix. Thus, the death rate from the population is  $1734/271822 = 0.00638$

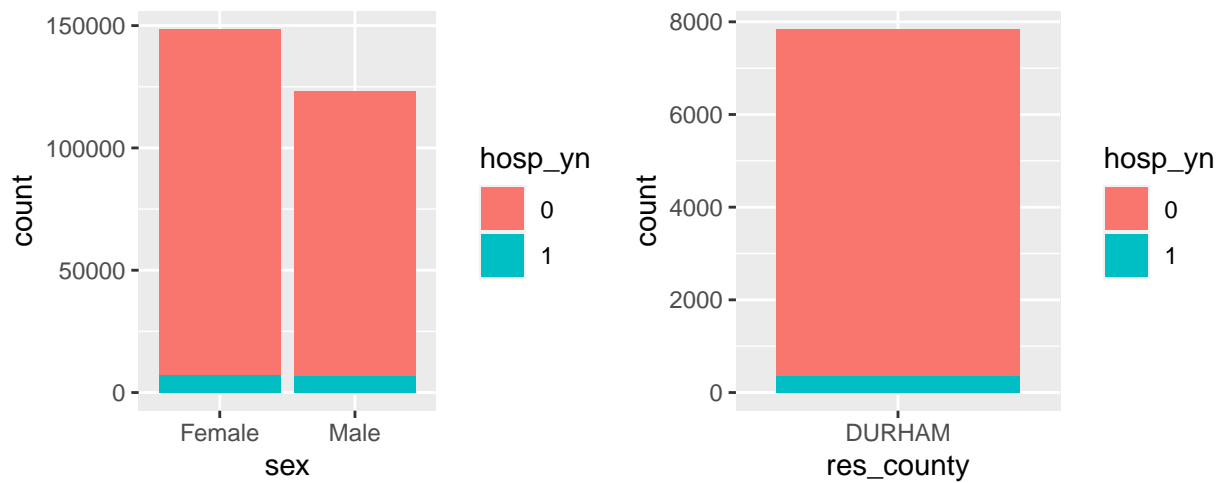
## Q7 Any finding during the EDA?



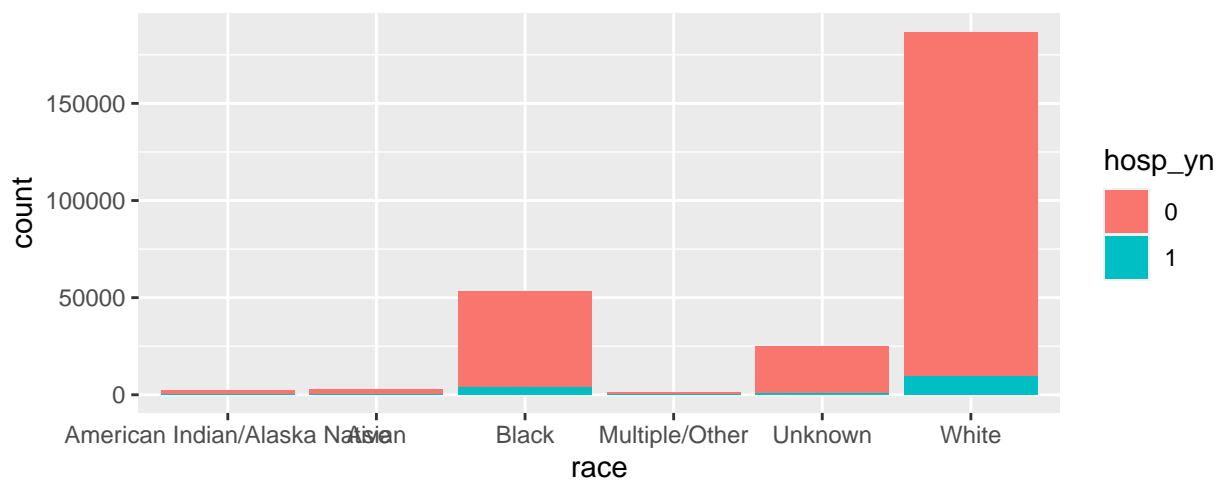
From this stacked bar chart we can observe the surge of active cases and death cases in December 2020.

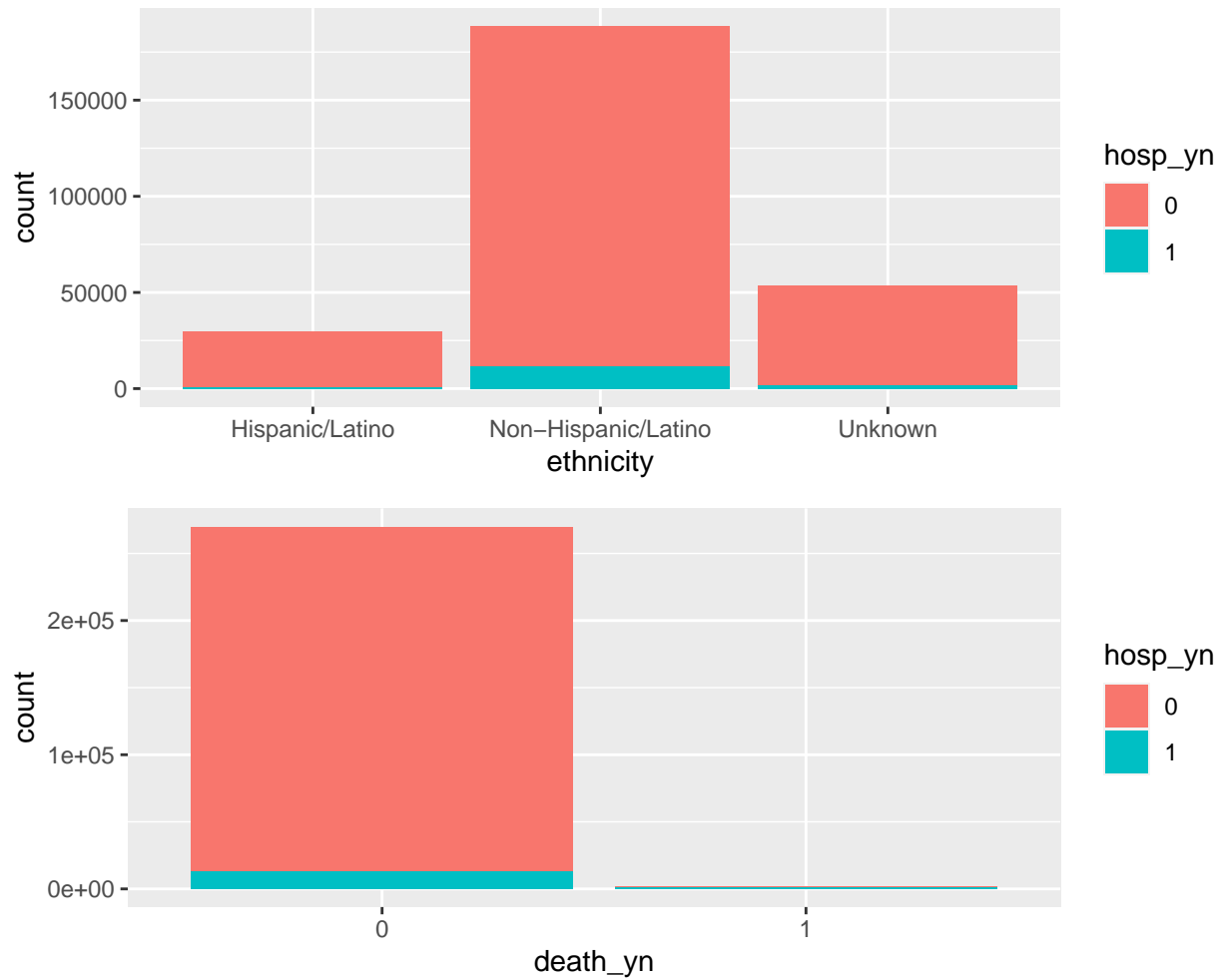


Most of the cases are people who are 18-49 years old, while the group whose ages are above 65 years old tends to have the highest number of death cases.



There is not a clear gender effect on the mortality of COVID based on this stacked bar chart.

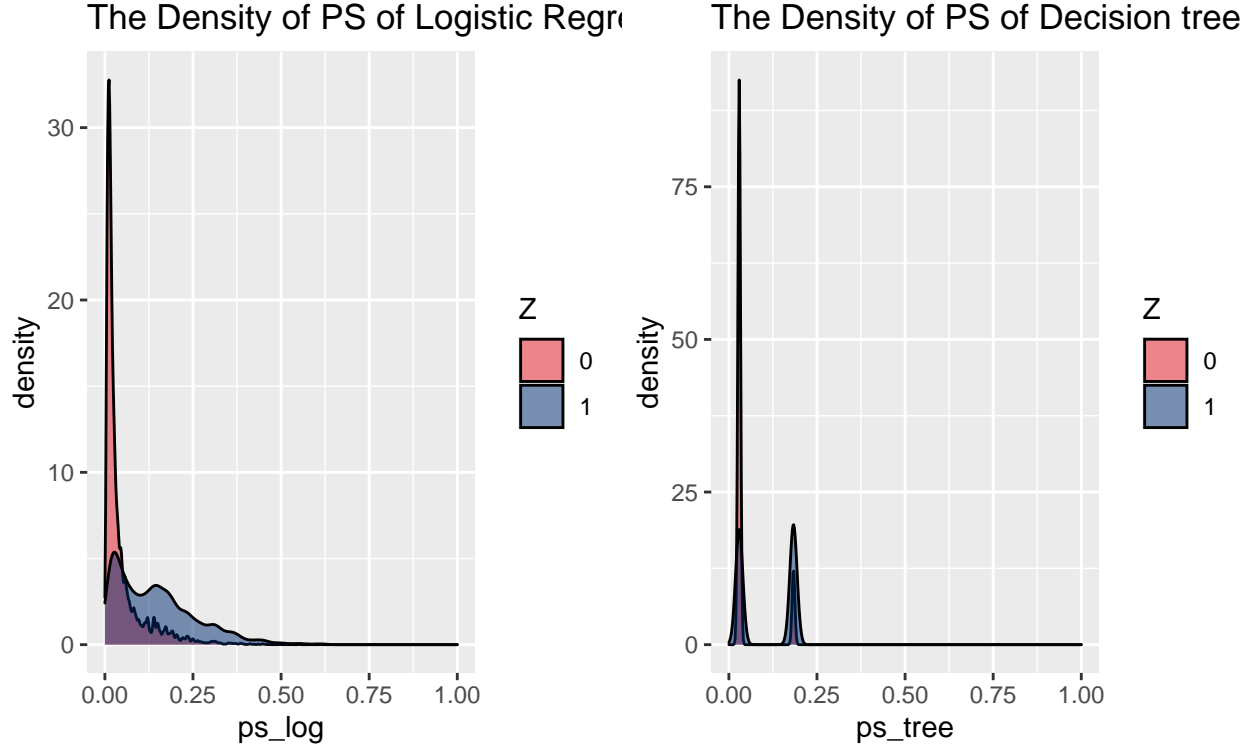




Most of the observations in this data set are Caucasians. Such imbalance may downplay the potential racial effect. Most of the death cases were hospitalized.

#### Q8 How do you build the model and what is your result?

For both outcome model and PS model, we apply two algorithms: Logistic Regression and Decision Tree. We want to compare the result of all four combinations.



The density plot of PS in different algorithms is provided to check the overlap. As we can see, both PS models have relative nice overlaps. So there is no need to clean any data out.

Table 1: The four different estimator treatment effect of Covid death rate

$\tau_1$	$\tau_2$	$\tau_3$	$\tau_4$
0.0170229	0.0172924	0.0178099	0.0181142

The table here illustrate our result, where:

- $\tau_1$  applied the Logistic Regression in both PS model and outcome model
- $\tau_1$  applied the Logistic Regression in outcome model and Decision Tree in PS model
- $\tau_1$  applied the Logistic Regression in PS model and Decision Tree in outcome model
- $\tau_4$  applied the Decision Tree in both PS model and outcome model

#### Q9 How Do you interpretate your reuslt?

All of the four estimators illustrate that the  $\tau^{DR}$  is about 0.018, which means the death rate of people sent to hospitals is about 0.018 higher than that of dying people who are not. Compared to the average death rate 0.0063, it is a very significant difference. This coefficient seems reliable since we can assume patients at hospitals would have more severe symptoms. Also, since we have no covariates describing each observation's severity of symptoms, such information may contribute to the difference in death rates. Moreover, the four estimators present close results, consistent with Bang and Robin's (2005) idea about how the Double Robust Estimator works. However, we can not exclude the possibility of random probability effects since our data provides a low death rate.

### **Q10: Are there any past literature? What are the differences between this project and them?**

In the past two years, several papers have discussed the potential determinants of Covid-19 death rates. Ian Feinhandler and four other authors offer several predictors that lead to the death rate during the first eight months of 2020. They implement the OLS model/Two-stage regression model/Lasso regression model and conclude that the national Covid-19 death rate is greater than that of other flu pandemics. Also, the increase in the reported death rate in states with Democratic governors is higher than the increase in states with Republican governors. (Feinhandler et al., 2020). Besides, in the paper Determinants of COVID-19 Death Rate in Europe: Empirical Analysis, six authors use the OLS models to test multiple hypotheses. They finally prove that the population density in European countries does not affect the COVID-19 death rate. Also, the COVID-19 death rate will not drastically raise mortality statistics since people already at risk are susceptible to the disease. (Kozlovskiy et al., 2021)

Comparing to other method, we focus on the 2020 NC data, and we choose to build the model under the causal inference, which allow us to explore the effect of hospitalization on death rate without the effects of other parameter. If we use a regression, for example, directly, it is hard for us to address the correlation variance between the death rate and the hospitalization is because of themselves, or because they are just both correlated to any other parameter.

### **Q11 Are there any limitations?**

The first limitation is that although we have many observations with solid values, there are still observations with missing values up to twenty thousand that we have to delete. As a result, there might exist a loss of “explainability” inside the observation with missing values of death rates. Besides, as mentioned above, our data has no access to the severity of symptoms the patient experiences. Thus, when testing the causal relationship between hospitalization and death rates, we can not exclude the possibility that patients with severer symptoms would go to hospitals and patients with less severe symptoms would stay home. This hypothesis, if true, would lead to the case that death rates of the former would be higher than that of the latter.

### **Q12 Are there any future directions/Perspective?**

With access to data regarding the severity of the illness, we can further test the selection bias mentioned above and future test the effectiveness of hospitalization and therapy received inside. Also, suppose we can acquire data related to the severity of other illnesses which share similar medical-source-occupation patterns with Covid-19. In that case, we are able to determine whether the virus is as severe as we expect.

## **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>

Robins, J. M., Rotnitzky, A., Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American statistical Association*, 89(427), 846-866

Feinhandler, Ian, et al. “Predictors of Death Rate during the Covid-19 Pandemic.” *Healthcare*, vol. 8, no. 3, 2020, p. 339., <https://doi.org/10.3390/healthcare8030339>.

Kozlovskiy, Serhii, i in. „Determinants of COVID-19 Death Rate in Europe: Empirical Analysis”. *Problemy Ekorożwoju*, t. 16, nr 1, 1, Polska Akademia Nauk. Komitet Człowiek i Środowisko PAN, 2021, s. 17–28.

Knittel, Christopher, and Bora Ozaltun. “What Does and Does Not Correlate with Covid-19 Death Rates.” NBER, 2020, <https://doi.org/10.3386/w27391>.

Centers for Disease Control and Prevention. “COVID-19 Case Surveillance Public Use Data.” 7 Apr. 2022, <https://data.cdc.gov/Case-Surveillance/COVID-19-Case-Surveillance-Public-Use-Data-with-Ge/n8mc-b4w4/data>. Accessed 10 Apr. 2022.



## 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
## 2020-12:74139  OTHER      : 38023  0 - 17 years : 25183  Female:148469
## 2020-11:56715  MECKLENBURG: 33373  18 to 49 years:154162  Male :123353
## 2020-10:35374  WAKE      : 19589  50 to 64 years: 53730
## 2020-07:28173  GUILFORD  : 11579  65+ years   : 38747
## 2020-08:22296  CUMBERLAND : 8979
## 2020-06:21578  FORSYTH   : 8484
## (Other):33547  (Other)    :151795
##
##      race      ethnicity
## American Indian/Alaska Native: 2334  Hispanic/Latino : 29868
## Asian      : 2772  Non-Hispanic/Latino:188461
## Black      : 53636  Unknown         : 53493
## Multiple/Other      : 1086
## Unknown      : 25048
## White      :186946
##
##      hosp_yn      death_yn      ps_log      ps_tree
## Min. :0.00000  Min. :0.000000  Min. :0.001228  Min. :0.02927
## 1st Qu.:0.00000  1st Qu.:0.000000  1st Qu.:0.012096  1st Qu.:0.02927
## Median :0.00000  Median :0.000000  Median :0.022454  Median :0.02927
## Mean :0.05122  Mean :0.006379  Mean :0.051221  Mean :0.05122
## 3rd Qu.:0.00000  3rd Qu.:0.000000  3rd Qu.:0.057002  3rd Qu.:0.02927
## Max. :1.00000  Max. :1.000000  Max. :0.615006  Max. :0.18327
##
```

## The number of death and survivor for the cleaning data

```
## # A tibble: 2 x 2
##   death_yn 'n()'
##   <dbl> <int>
## 1      0 270088
## 2      1  1734
```

## 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)
library(tree)
library(rpart)
library(gridExtra)
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()%>%
  filter(hosp_yn!="Unknown")%>%
  mutate(death_yn=ifelse(death_yn=="Yes",1,0),
         hosp_yn=ifelse(hosp_yn=="Yes",1,0))%>%
  filter(sex!="Missing")
county_list=covid_clean%>%
  group_by(res_county)%>%
  summarise(observation=n())
covid_clean=covid_clean%>%
  mutate(res_county=ifelse(res_county %in% county_list$res_county[county_list$observation < 2000], "OTHER", res_county))
covid_clean=as.data.frame(unclass(covid_clean),stringsAsFactors=TRUE)
stackedbar_case_month_hosp_yn=ggplot(covid_clean, aes(x = as.factor(case_month),
              fill = as.factor(hosp_yn))) +
  geom_bar(position = "stack") +
  guides(fill=guide_legend(title="hosp_yn")) +
  scale_x_discrete(name="case_month")
stackedbar_res_county_hosp_yn=ggplot(covid_clean[covid_clean$res_county=="DURHAM",], aes(x = as.factor(res_county),
              fill = as.factor(hosp_yn))) +
  geom_bar(position = "stack") +
  guides(fill=guide_legend(title="hosp_yn")) +
  scale_x_discrete(name="res_county")
stackedbar_age_group_hosp_yn=ggplot(covid_clean, aes(x = as.factor(age_group),
              fill = as.factor(hosp_yn))) +
  geom_bar(position = "stack") +
  guides(fill=guide_legend(title="hosp_yn")) +
  scale_x_discrete(name="age_group")
stackedbar_sex_hosp_yn=ggplot(covid_clean, aes(x = as.factor(sex),
              fill = as.factor(hosp_yn))) +
  geom_bar(position = "stack") +
  guides(fill=guide_legend(title="hosp_yn")) +
  scale_x_discrete(name="sex")
stackedbar_race_hosp_yn=ggplot(covid_clean, aes(x = as.factor(race),
              fill = as.factor(hosp_yn))) +
  geom_bar(position = "stack") +
  guides(fill=guide_legend(title="hosp_yn")) +
  scale_x_discrete(name="race")
stackedbar_ethnicity_hosp_yn=ggplot(covid_clean, aes(x = as.factor(ethnicity),
              fill = as.factor(hosp_yn))) +

```

```

    geom_bar(position = "stack") +
    guides(fill=guide_legend(title="hosp_yn")) +
    scale_x_discrete(name="ethnicity")
stackedbar_death_yn_hosp_yn=ggplot(covid_clean, aes(x = as.factor(death_yn),
    fill = as.factor(hosp_yn))) +
    geom_bar(position = "stack") +
    guides(fill=guide_legend(title="hosp_yn")) +
    scale_x_discrete(name="death_yn")
stackedbar_case_month_hosp_yn
stackedbar_age_group_hosp_yn
grid.arrange(stackedbar_sex_hosp_yn, stackedbar_res_county_hosp_yn,
    ncol=2)
stackedbar_race_hosp_yn
stackedbar_ethnicity_hosp_yn
stackedbar_death_yn_hosp_yn
att=data.frame(tau1=numeric(1),tau2=numeric(1),tau3=numeric(1))
## get the PS
ps_model_log=glm(hosp_yn~.-death_yn,family="binomial",data=covid_clean)
ps_model_tree=rpart(hosp_yn~.-death_yn,method="anova",data=covid_clean)

covid_clean=covid_clean%>%
    mutate(ps_log=ps_model_log$fitted.values,
        ps_tree=predict(ps_model_tree,covid_clean))
q1=ggplot(covid_clean, mapping = aes(x = ps_log, fill = factor(hosp_yn)))+
    geom_density(alpha = .5) +
    scale_fill_manual(values = c("#E81828", "#002D72"))+
    xlim(c(0,1))+
    guides(fill=guide_legend(title="Z"))+
    ggtitle("The Density of PS of Logistic Regression for different group ")
q2=ggplot(covid_clean, mapping = aes(x = ps_tree, fill = factor(hosp_yn)))+
    geom_density(alpha = .5) +
    scale_fill_manual(values = c("#E81828", "#002D72"))+
    xlim(c(0,1))+
    guides(fill=guide_legend(title="Z"))+
    ggtitle("The Density of PS of Decision tree for different group ")
grid.arrange(q1, q2, ncol=2)
##seperate the data
covid_clean_1=covid_clean%>%
    filter(hosp_yn==1)
covid_clean_0=covid_clean%>%
    filter(hosp_yn==0)
## generate the outcome model
m1_log=glm(death_yn~.-hosp_yn-ps_log-ps_tree,family = "binomial",
    data=covid_clean_1)
m0_log=glm(death_yn~.-hosp_yn-ps_log-ps_tree,family = "binomial",
    data=covid_clean_0)

m1_tree = rpart(death_yn~.-hosp_yn-ps_log-ps_tree,
    data=covid_clean_1)
m0_tree = rpart(death_yn~.-hosp_yn-ps_log-ps_tree,
    data=covid_clean_0)

## predict m1,m1

```

```

## calculate the result
result=covid_clean%>%
  mutate(m1_log=predict.glm(m1_log,covid_clean,type = "response"),
         m0_log=predict.glm(m0_log,covid_clean,type = "response"),
         m1_tree=predict(m1_tree,covid_clean),
         m0_tree=predict(m0_tree,covid_clean))%>%
  mutate(tau1=m1_log+hosp_yn*(death_yn-m1_log)/ps_log-m0_log-(1-hosp_yn)*(death_yn-m0_log)/(1-ps_log),
         tau2=m1_log+hosp_yn*(death_yn-m1_log)/ps_tree-m0_log-(1-hosp_yn)*(death_yn-m0_log)/(1-ps_tree),
         tau3=m1_tree+hosp_yn*(death_yn-m1_tree)/ps_log-m0_tree-(1-hosp_yn)*(death_yn-m0_tree)/(1-ps_log),
         tau4=m1_tree+hosp_yn*(death_yn-m1_tree)/ps_tree-m0_tree-(1-hosp_yn)*(death_yn-m0_tree)/(1-ps_t.
  summarise(tau1=mean(tau1),
            tau2=mean(tau2),
            tau3=mean(tau3),
            tau4=mean(tau4))
result%>%
  kable(caption = "The four different estimator treatment effect of Covid death rate",
        col.names = c("$\\tau_1$",
                      "$\\tau_2$",
                      "$\\tau_3$",
                      "$\\tau_4$"),
        align = "c", booktabs = TRUE, escape = FALSE)%>%
  kable_styling(latex_options = "HOLD_position")

summary(factor_covid)
summary(covid_clean)
covid_clean%>%
  group_by(death_yn)%>%
  summarise(n())

```